DOCUMENT RESUME

ED 437 439 TM 030 608

TITLE	Peer Reviewer Guidance for Evaluating Evidence of Final Assessments under Title I of the Elementary and Secondary Education Act.
INSTITUTION	Department of Education, Washington, DC.
PUB DATE	1999-11-00
NOTE	99p.
PUB TYPE	Reports - Descriptive (141)
EDRS PRICE	MF01/PC04 Plus Postage.
DESCRIPTORS	*Academic Standards; Accountability; *Compensatory Education; Educational Assessment; Elementary Secondary Education; *Equal Education; *Evaluation Methods; *Peer Evaluation; *State Programs
IDENTIFIERS	*Elementary Secondary Education Act Title I

ABSTRACT

The reauthorization of the Elementary and Secondary Education Act in 1994 includes explicit requirements to ensure that students served by its Title I are given the same opportunity to achieve high standards and are held to the same high expectations as all students in each state. By the spring of 2000, states should be prepared to submit evidence that final assessment systems are in place. To determine whether states have met Title I requirements, the U.S. Department of Education will use a peer review process involving experts in the fields of standards and assessments. This document is intended both to inform states about useful evidence and to guide teams of peer reviewers. The document consists of three main sections: (1) "General Characteristics of the Assessment System"; (2) "The Core of the Assessment System"; and (3) "Reporting and Using Assessment Results in Accountability." Subsections deal with specific aspects of these broad areas, such as "meeting standards of technical quality." Most subsections include discussions of the requirements and their intent and purpose and questions peer reviewers should ask. Four appendixes clarify a requirement related to students of limited English proficiency, discuss assessment information flow, and present a summary of types and sources of evidence for the reviews. (SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

TM

ED 437 439

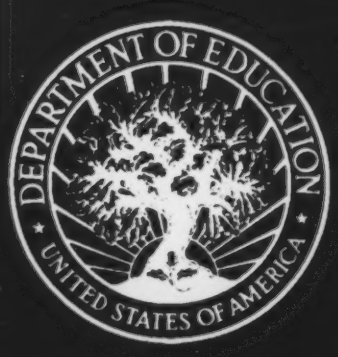
PEER REVIEWER GUIDANCE

FOR EVALUATING EVIDENCE OF FINAL ASSESSMENTS UNDER TITLE I OF THE ELEMENTARY AND SECONDARY EDUCATION ACT

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

- [x] This document has been reproduced as received from the person or organization originating it.
- [ ] Minor changes have been made to improve reproduction quality.

- Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.



United States Department of Education

November 1999

TM030608

2

BEST COPY AVAILABLE

PEER REVIEWER GUIDANCE FOR EVALUATING EVIDENCE OF FINAL ASSESSMENTS UNDER TITLE I OF THE ELEMENTARY AND SECONDARY EDUCATION ACT

Contents

Introduction			2
Part I.	General Characteristics of the Assessment System		
	A:	Content, Grade Levels, and Administration	6
	B:	Inclusion	12
Part II.	The Core of the Assessment System		
	C.	Assessments Must be Aligned to Standards	22
	D.	Meeting Professional Standards of Technical Quality	31
Part III.	Reporting and Using Assessment Results in Accountability		
	E.	Providing Individual Reports	46
	F.	Disaggregated Reporting	48
	G.	Development of District and School Profiles	50
	H.	Ensuring that State Assessments are the Primary Basis for Determining LEA and School Progress	52
	I.	Include Students who have Attended School in the LEA for a Full Academic Year	55
Appendix A		Including LEP Students in State Assessments under Title I: "To the extent practicable"	59
Appendix B		Must All the Standards Be Assessed?	69
Appendix C		Summary of Alignment Elements and Illustrative Types and Sources of Evidence	74
Appendix D		Summary of Elements and Illustrative Evidence of Technical Quality	78

INTRODUCTION

Raising academic standards for all students and measuring student performance to hold schools accountable for educational progress are central strategies for promoting educational excellence and equity in our schools. The reauthorization of the Elementary and Secondary Education Act in 1994 reformed federal programs to support State efforts to establish challenging standards, to develop aligned assessments, and to build accountability systems for districts and schools that are based on educational results. In particular, the Act includes explicit requirements to ensure that students served by Title I are given the same opportunity to achieve to high standards and are held to the same high expectations as all students in each State.

Title I required States to adopt or develop challenging content and performance standards by the 1997-98 school year. It also requires States to develop and implement assessments aligned to those standards and accountability systems based on student performance against those standards by the 2000-01 school year. Assessments must be field-tested prior to implementation. Thus, by the spring or summer of 2000, States should be prepared to submit evidence that their final assessment systems are in place.

The purpose of this guidance is twofold: 1) to inform States what would be useful evidence to demonstrate that they have met Title I final assessment requirements; and 2) to guide teams of peer reviewers who will examine evidence submitted by States and advise the Department on whether a State has met Title I requirements. The intent of these requirements is to help States develop comprehensive assessment systems that provide accurate and valid information for holding districts and schools accountable for student performance against State standards. Although this document addresses each requirement separately, reviewers and States should recognize that the requirements are interrelated and that decisions about whether a State has met the requirements will be based on comprehensive examination of the evidence submitted.

The Peer Review Process

To determine whether States have met Title I assessment requirements, the U.S. Department of Education will use a peer review process involving experts in the fields of standards and assessments. The review will evaluate State assessment systems against Title I requirements only. In other words, reviewers will examine characteristics of State assessment systems that will be used to hold schools and school districts accountable under Title I. They will not assess compliance of State assessment systems with other Federal laws such as Title VI of the Civil Rights Act of 1964, Section 504 of the Rehabilitation Act of 1973, or provisions of the Individuals with Disabilities Education Act. The fact that an assessment system meets Title I assessment requirements does not necessarily mean that it complies with other laws. For guidance on compliance with Federal civil rights laws, States may consult with the Department of Education's Office for Civil Rights.

Furthermore, the peer review process will not directly examine a State's assessment instruments or specific test items. Rather, it will examine *evidence* compiled and submitted by each State that is intended to show that its assessment system meets Title I requirements. Such evidence may include, but is not limited to, results from alignment studies; results from validation studies; written policies on including and, if appropriate, providing accommodations for students with disabilities and limited English proficient students; written policies on native-language testing of

LEP students; and score reports showing disaggregation of student performance data by statutorily specified categories. Peer reviewers will advise the Department on whether a State assessment system meets a particular requirement based on the totality of evidence submitted. They will also provide constructive feedback to help States strengthen their assessment systems.

States are invited to submit evidence of Title I compliance as soon as they have adopted or developed their final assessment systems. The Department will conduct peer reviews of submissions received by the beginning of each quarter (January 1, April 1, July 1, and October 1) during that quarter.

Statutory and Regulatory Requirements for Final Assessment Systems

Each State must adopt or develop a final assessment system aligned to State content and performance standards by the beginning of the 2000-01 school year, and tests must be administered before the end of the 2000-01 school year. Although most States are developing statewide assessment systems applicable to all students in the State, note that Title I assessment requirements also apply to States that, instead of developing a statewide system, choose to develop an assessment system applicable only to students served by Title I.

Title I requires State assessment systems to have the following characteristics:

- Assessments must be aligned with State content and performance standards, and they must provide coherent information about student attainment of State standards in at least math and reading/language arts.
- If the State measures the performance of all children, the same assessments must be used to measure the performance of students served by Title I.
- Assessments must be administered annually to students in at least one grade in each of three grade ranges—grades 3 through 5, grades 6 through 9, and grades 10 through 12.
- The assessment system must provide for
  - ◊ participation in the assessments of all students in the grades being assessed;
  - ◊ reasonable adaptations and appropriate accommodations for students with diverse learning needs, where such adaptations or accommodations are necessary to measure the achievement of those students relative to State standards; and
  - ◊ inclusion of LEP students, who shall be assessed, to the extent practicable, in the language and form most likely to yield accurate and reliable information on what they know and can do to determine their mastery of skills in subjects other than English. To meet this requirement, States shall make every effort to use or develop linguistically accessible assessment measures, and they may request assistance from the Secretary if those measures are needed.
- The assessment system must involve multiple approaches with up-to-date measures of student performance, including measures that assess complex thinking skills and understanding of challenging content.
- Assessments must be used for purposes for which they are valid and reliable, and they must meet relevant, nationally recognized, professional and technical standards for quality. A State may include assessment measures that do not meet these requirements as one of

multiple measures if it provides sufficient information regarding its efforts to validate the measures and to report the results of those validation studies.

- Assessment results must be disaggregated within each school and district by gender, major racial and ethnic groups, English proficiency status, migrant status, students with disabilities as compared to students without disabilities, and economically disadvantaged students as compared to students who are not economically disadvantaged. Disaggregated data must be included in annual school profiles.
- The assessment system must provide individual student interpretive and descriptive reports that include individual scores or other information on the attainment of student performance standards.

A State may request a one-year extension from the Secretary if it finds problems during field-testing and submits a strategy for correcting those problems. If a State has not developed or adopted a standards-based assessment system that measures performance in at least math and reading/language arts by the 2000-01 school year, and if an extension is denied, the State must adopt an assessment system that meets Title I requirements, such as a system adopted by another State and approved by the Department, if appropriate.

Starting in the 2000-01 school year, the statewide assessment system will be the primary means for determining whether schools and school districts receiving Title I funds are making adequate progress toward educating students to high standards. In determining the progress of schools, States must include scores of all students assessed who have attended the school for at least a full academic year. In determining the progress of school districts, States must include scores of students who have attended school in the district for a full academic year, even if they have attended multiple schools.

Because Title I makes State assessment systems central to holding schools and districts accountable, this document focuses on the uses of State assessment systems at the school and district levels. Nevertheless, peer reviewers should note that the Title I requirements listed above include the requirement that State assessment systems report results at the level of individual students.

State and Local Roles

Roles and responsibilities within a State assessment system are allocated at the State, district, and school levels. The Department's 1997 guidance, *Standards, Assessment, and Accountability*, describes three acceptable state-local configurations for final assessment systems:

- The *state* model, in which all students are assessed with a common State instrument that yields data for determining adequate yearly progress for all schools and school districts;
- The *mixed* model, in which State assessments are supplemented by State-approved local assessments; and
- The *local* model, in which the State uses no common instrument and instead applies uniform standards to approve and monitor assessment systems developed by each district.

In implementing final assessment systems, States have two main responsibilities: 1) They must develop, score, and report findings from State assessments, and 2) they must promulgate rules

and procedures for local assessment systems, as well as monitor such systems, to ensure technical quality and compliance with Title I requirements. The second function is particularly significant in assessment systems with strong local responsibility. Yet it remains salient even for States with uniform statewide assessments, since many such States employ a mixed model with local assessments playing some role in meeting Title I requirements.

Format of the Guidance

This document consists of three main sections with several subsections:

I. General Characteristics of the Assessment System
   - A. Content, Grade Levels, and Administration
   - B. Inclusion

II. The Core of the Assessment System
   - C. Assessments Must Be Aligned to Standards
   - D. Meeting Professional Standards of Technical Quality

III. Reporting and Using Assessment Results in Accountability
   - E. Providing Individual Reports
   - F. Disaggregated Reporting
   - G. Development of District and School Profiles
   - H. Ensuring that State Assessments Are the Primary Basis for Determining LEA and School Progress
   - I. Include Students Who Have Attended School in the LEA for a Full Academic Year

Most subsections include five parts:

1. **Requirements:** Statutory and regulatory excerpts
2. **Intent and purpose:** A brief discussion of the reasoning behind the requirement
3. **Abbreviated description:** A description of critical points in the requirement
4. **Full description:** Detailed explanation of each point in the requirement
5. **Questions for reviewers:** Questions peer reviewers will consider as they look at State evidence, accompanied by examples of "desirable evidence" that States might provide for each requirement as well as evidence likely to be considered "incomplete" or "unacceptable."

The document also includes four appendices. Appendix A clarifies the requirement that States assess LEP students "to the extent practicable" in the language and form most likely to yield accurate and reliable information on what these students know and can do in subjects other than English. Appendix B discusses the flow and possible uses of assessment information gathered at State and local levels. Appendix C discusses the types of evidence useful for demonstrating alignment between standards and assessments. Appendix D discusses the types of evidence useful for demonstrating technical quality.

PART I: GENERAL CHARACTERISTICS

Part IA. Content, Grade Levels, and Administration

1. Requirement

> Each State plan shall demonstrate that the State has developed or adopted a set of high-quality, yearly student assessments, including assessments in at least mathematics and reading or language arts, that will be used as the primary means of determining the yearly performance of each local educational agency and school served under this part in enabling all students to meet the State's student performance standards. Such assessments shall –
>
> - be the same assessments used to measure the performance of all children, if the State measures the performance of all children;[1]
> - measure the proficiency of students in the academic subjects in which a State has adopted challenging content and student performance standards and be administered at some time during grades 3 through 5, grades 6 through 9, and grades 10 through 12.
> - involve multiple up-to-date measures of student performance, including measures that assess higher order thinking skills and understanding (Sec. 1111(b)(3)(A), (D), and (E)).

2. Intent and purpose

The intent of these requirements is to ensure that 1) Title I students are not held to lower standards than other students through less rigorous assessments, or through assessments that measure different standards; and 2) schools and districts know how well all of their students are doing in relation to a common set of State standards so that schools and districts can be held accountable and make improvements.

The requirement for including multiple measures has several purposes:

1) to provide more complete measurement of the content and performance standards and therefore increase the validity of the inferences made about school performance;
2) to offer a variety of opportunities for schools and districts to demonstrate performance and therefore increase the fairness of determinations about performance;
3) to provide a means for the State assessment to measure a range of cognitive attributes, including higher order thinking skills; and
4) to ensure that the State standards are assessed comprehensively and with the same degree of emphasis and depth as stated in the standards.

States may meet these requirements by implementing statewide tests, local tests that are approved by the State, or both. Most States are implementing State assessment *systems* that include measures at the State, district, and school levels. Describing such a system and how the

[1] Questions regarding modifications and adaptations for students with diverse learning needs and linguistically accessible assessments for limited English proficient students are addressed in Part I B, "Inclusion."

various components work together is very important for helping peer reviewers examine State evidence in relation to the Title I requirements.

3. Full description

The State assessment system may consist of standards-based measures adopted or developed by the State, measures adopted or developed by LEAs, or both. If determination of school and district progress is based in whole or in part on measures adopted/developed by LEAs, the State must provide criteria or models for the LEA assessments. The State also is responsible for monitoring the quality of such assessments.

Content[2]

The law requires the annual assessment of all students served by Title I in both mathematics and reading or language arts, using measures that assess higher order thinking skills and understanding. Although the law requires only that States assess in these content areas, the regulations make it clear that the Secretary of Education encourages States to broaden their assessments to include other content areas such as science and social studies: "If a State has standards and assessments for all students in subjects beyond mathematics and reading/language arts, the regulations do not preclude a State from including, for accountability purposes, additional subject areas, and the Secretary encourages them to do so." (Federal Register, July 3, 1995, Regulations, page 34800)

If the State assesses mathematics or reading/language arts through testing in another content area, the assessments must yield information about student performance in mathematics and reading/language arts. If a State assesses math and reading in the same grade using a matrix sampling approach, it must ensure that every student is assessed in both reading and math. Also, if a language arts assessment is used, it must measure reading and yield information about student performance in reading.

Multiple Measures[3]

The Title I legislation requires the use of multiple measures. This requirement has been interpreted in the Department's *Guidance on Standards, Assessments, and Accountability* (1997) to mean that different approaches and formats should be included in a State assessment system. Examples include criterion-referenced tests, standardized norm-referenced tests, writing samples, completion of graphic representations, observation checklists, performance of exemplary tasks, performance events, and portfolios of student work. The assessments must include measurement of complex skills and understanding of challenging content in at least mathematics and reading/language arts.

Although multiple approaches such as a criterion-referenced test and a performance task for a single subject are not required, States should determine how multiple approaches are appropriately included in their State assessment systems based on 1) the nature of the content

[2] The degree to which content standards in mathematics and reading/language arts are assessed and the quality of coverage are addressed in Part II C, "Assessments Aligned to Standards."

[3] The quality and use of multiple measures is addressed in Part II C, "Assessments Aligned to Standards," and Part II D, "Professional Standards of Technical Quality."

and performance standards in each content area and 2) the contribution of the measures to the technical quality of the assessment system at the level of use of the results. For example, depending upon the findings of the State's alignment study, multiple measures may be necessary in order to adequately assess the State's standards and therefore meet the alignment requirement. Multiple approaches may also provide more complete information on school progress.

For purposes of holding districts and schools accountable for making adequate yearly progress, the State assessment must be the primary measure used. The State may use data derived from other school indicators such as attendance and graduation. However, such data may not be used to meet the Title I requirement for multiple measures in a statewide assessment system.

Grade Levels

Testing only needs to occur in one grade of each gradespan, and although at least mathematics and reading/language arts must be assessed within each grade span, different subjects may be tested in different grades (e.g., reading in grade 3 and math in grade 4). The performance of each school served by Title I, however, must be measured. If a school does not encompass a grade that is within the State assessment system (e.g., a school serves grades k-2 and testing is in grade 4), then the State must develop a means for holding such schools accountable for student performance. Such schools may either 1) use locally adopted or developed assessments or 2) use assessment information from their receiving schools that shows student performance in math and reading within the relevant grade spans.

Administration

If the State measures the performance of all students, these assessments must be used to measure the progress of students in Title I schools. The State should provide a description of its statewide assessment system that explains the purpose of its State assessment and how it is administered throughout the State.

4. Preparing a State submission of evidence

A State should submit a narrative description of its assessment system that explains the purpose of the system, the various components of the system, the subjects and grades assessed, and how assessment results are used. State submissions should address each of the peer review questions listed in the next section.

A narrative description supported by other relevant documents would be most helpful. For example:

- sample score reports would provide information about what is assessed and how content and performance standards are communicated;
- assessment development materials such as test blueprints and item specifications would demonstrate that multiple measures are used and that a range of cognitive attributes are considered in the assessment system;
- criteria or models that are provided to LEAs for adoption or development of local assessments would help explain how the State ensures quality in local components of its system; and

- administration manuals that explain allowable accommodations for students with disabilities and LEP students would support a description of a State assessment system that includes all students.

5. Questions for peer reviewers

Peer reviewer questions	Desirable Evidence	Incomplete or Unacceptable Evidence
A1. Does the State have a statewide system for assessing all schools in the selected grade spans, including Title I schools? If not, does the State at least have a system for assessing students in Title I schools in relation to performance on State standards?	Reviewers will determine whether the State assessment system (or, if there is no State system, the assessment used for Title I purposes) includes all required content areas and grade levels, which students are covered by the system, and information about how results are used.	The State uses a different system of assessment for some groups of students, such as Title I students. The State uses a different system of assessment to measure achievement in Title I schools than it uses to measure the achievement in Title I schools. The State has no provisions for measuring achievement in Title I schools.
A2. Does the State assessment system measure the performance of students in Title I schools using a statewide test, local assessments, or some combination? If the State assessment system includes LEA-adopted or developed assessments, how does the State ensure the quality and rigor of the assessments?	If local assessments are used, peer reviewers will look for evidence that the State has a means of ensuring high quality and rigor in local assessments. Evidence of monitoring the quality and use of local assessments might include State-provided criteria or models for local assessments, a peer review process for local assessments, or other procedures to monitor the quality of the assessments and their administration and use.	LEAs choose their assessments with no oversight (either models of assessments or criteria for selection) from the State. The State does not monitor the quality or rigor of local assessments.
A3. How does the State evaluate the effectiveness of schools that do not contain any of the grade spans covered by the State assessment system (e.g., k-2 schools)?	Reviewers will look for evidence such as descriptions of LEA assessments or a method for matching scores from receiving schools.	The State does not collect achievement data in schools with grades outside the required grade spans and has not developed methods for evaluating the effectiveness of Title I schools that do not contain the required grade spans.

Peer reviewer questions	Desirable Evidence	Incomplete or Unacceptable Evidence
A4. How does the State incorporate multiple measures of student achievement?	Reviewers will look for a description of multiple approaches or instruments, based on State content and performance standards, in the State system. Examples might include the use of multiple approaches and formats within a single test; the use of multiple assessment instruments; and the use of a writing test as well as a reading test This evidence might be found in descriptions of the assessment system, test blueprints, or item specifications. Evidence of the use of local assessments to measure content areas or standards will also be considered by reviewers.	The State uses only a multiple-choice test.
A5. Are the assessments administered annually, covering the required grade spans and content areas, incorporating the measurement of higher order thinking skills and understanding, and yielding scores in at least mathematics and reading?	Reviewers will look for evidence such as that needed to fill in the chart below. This evidence may come from documents such as score reports, State reports of assessment results, test blueprints, or descriptions of the State assessment program. If the State uses assessments in content areas other than mathematics and reading/language arts to assess proficiency in mathematics and reading/language arts, reviewers will look for evidence of the production of (sub) scores in at least reading and math.	The State measures only basic skills in reading/language arts or mathematics.

Evidence of Required Assessments, by Subject and Grade Span

	Grade Span 3-5	Grade Span 4-8	Grade Span 9-12
Administered annually			
Mathematics, including measurement of higher order thinking			
Reading/language arts, including measurement of higher order thinking			
Other subjects optional (specify)			
Scores reported in reading			
Scores reported in math			

Part IB. Inclusion

1. Requirement

The State assessments shall provide for –

- the participation in such assessments of all students in the grades being assessed;
- the reasonable adaptations and accommodations for students with diverse learning needs, necessary to measure the achievement of such students relative to State content standards; and
- the inclusion of limited English proficient students who shall be assessed, to the extent practicable, in the language and form most likely to yield accurate and reliable information on what such students know and can do, to determine such students' mastery of skills in subjects other than English.
  (Sec. 1111(b)(3)(F))
- The State plan shall identify the languages other than English that are present in the participating student population and indicate the languages for which yearly student assessments are not available and are needed. The State shall make every effort to develop such assessments and may request assistance from the Secretary if linguistically accessible measures are needed.
  (Sec. 1111(b)(5))

2. Intent and purpose

The purposes of these requirements are 1) to ensure that all students are held to the same high standards and appropriately assessed against those standards; and 2) to ensure that the indicators used to hold schools accountable include performance data on all students in the grades being assessed. States are responsible for assessing all such students relative to proficiency on the State's content and performance standards in mathematics and reading/language arts. States must show how they use a variety of strategies to make certain that all students participate in the assessment system. These strategies may include appropriate accommodations, alternate assessments, assessments in the students' primary languages, and linguistically simplified assessments.

3. Full description

States are responsible for assessing all students in the grades being assessed. Therefore, States must provide means to determine the achievement of students with disabilities and limited English proficient students relative to the State's content and performance standards when standard assessment procedures do not provide this information. This may be accomplished through providing appropriate accommodations in setting, scheduling, presentation, and response formats for the standard assessment, or through developing or adopting primary-language assessments or alternative assessment procedures tied to the content and performance standards.

The critical issue to consider in this section is whether the assessment system allows for assessing students with disabilities and limited English proficient students against the same content and performance standards that apply to all students. Technical quality, including

reliability and validity, must be ensured if assessments are administered in a non-standard manner.

Changes to Standard Assessment Procedures[4]

Accommodations are changes to standard assessment conditions, including changes in setting, scheduling, timing, presentation, and response. For best results, accommodations should be the same as the instructional conditions that the student normally experiences. Determining whether an accommodation compromises technical quality involves judging whether the accommodation either alters the construct assessed or changes the performance standard.

The best way to determine whether a specific accommodation produces a valid score (i.e., measures the same construct as that measured by the standard version) is through empirical research. Because conducting such research is time consuming and expensive, experts have developed several rules of thumb for categorizing some common changes to standard testing conditions (National Center on Educational Outcomes, 1997). For example, in most cases, providing a separate room for a student to take the test is considered an accommodation that produces a valid score. Allowing the use of a calculator on an assessment designed to measure calculation skills is usually considered to be an accommodation that produces an invalid score.

Students with Disabilities

Decisions on the types of assessment accommodations or adaptations provided to a student with disabilities, or the decision to use an alternate assessment, should follow standard State guidelines that are consistent with IDEA requirements. Decisions on each student's participation in State and district assessment programs must be consistent with the appropriate Federal laws and regulations. For some students with disabilities the appropriate law is the reauthorized 1997 Individuals with Disabilities Education Act (IDEA). Other students with disabilities who are evaluated and determined to be ineligible for special education and related services under IDEA are provided reasonable accommodations in accordance with Section 504 of the Rehabilitation Act of 1973 (Section 504), as amended.

Students with Individualized Education Programs (IEP) under IDEA: This Federal law and its accompanying regulations require that students with disabilities are included in State and district-wide assessment programs, with appropriate accommodations and modification, if necessary.

IDEA also recognizes that some students with disabilities may be unable to participate in general State or district assessments, even with the use of appropriate accommodations. States must ensure that guidelines are developed for the participation of students with disabilities in alternate assessments for those students who cannot participate in the general assessment program. (IDEA requires that states develop alternate assessments by July 1, 2000.)

A student's participation in alternate assessments or any individual assessment accommodations

[4] The terminology used for assessment alterations is confusing. The terms accommodations, modifications, adaptations, and alterations are sometimes used to mean the same thing, and sometimes used to mean different things. Because these terms are not used with uniform, consistent meaning, we only use the term "accommodation" here, with adjectives added to clarify whether an accommodation results in a valid score or invalid score.

or modifications needed by the student must be determined and documented in his or her IEP, by the student's IEP team. If a student's IEP team determines that he or she will not participate in a particular State or district assessment of student achievement (or part of an assessment), the student's IEP must include a statement of why that assessment is not appropriate for the student and how the student will be assessed. If IEP teams properly make individualized decisions about the participation of each child with a disability in State and district-wide assessments (including the use of appropriate accommodations and modifications in the administration, as appropriate), it should be necessary to use alternate assessments for only a relatively small percentage of children with disabilities (Federal Register, Vol. 64, No. 48, March 12, 1999, p.12564).

Students with 504 Plans or IEPs not Under IDEA: The student who meets the Section 504 definition of disability must be provided reasonable accommodations, which can include special education and related services if determined appropriate by the Section 504 placement team. Decisions on the types of assessment accommodations or adaptations provided to a student under Section 504 should be documented in the student's IEP (if the parent and school placement team have agreed upon the IEP option) or 504 Plan, and they should be closely related to procedures used in the student's instruction.

Alternate assessments are used when appropriate accommodations can not provide students with an opportunity to demonstrate their knowledge and skills. Typically, alternate assessments incorporate more fundamental changes to testing conditions, such as using an entirely different format for the assessment (e.g., a portfolio system instead of on-demand tests) or assessing content standards that are changed in some way (e.g., expanded standards with different performance descriptors).

Students with Limited English Proficiency

All LEP students in the grades being assessed must be a part of the State's assessment system. One of the pieces of the statute that is challenging to interpret, however, is what it means to assess LEP students "to the extent practicable in the language and form most likely to yield accurate and reliable results." The Department has developed a series of questions that a State or district should consider when determining the "extent practicable" for offering assessments in other languages and forms for LEP students. Appendix A includes this guidance and peer reviewers should consider how the State explains its system against this framework.

States must identify the languages other than English that are present in their student population and the levels of English proficiency among their LEP students, and use this information to determine if assessments written in languages other than English are needed. If a State has a large population of LEP students who speak a single non-English language, it may be feasible and appropriate to provide assessments aligned with standards in those languages. In most States, the population of Spanish-speaking students is large enough to justify the development of Spanish versions of the assessments. If several different languages are spoken by LEP students and no single language constitutes a significant concentration, it may not be feasible to assess in students' native languages, but appropriate accommodations should be offered that reflect the instructional approaches those students are experiencing.

At the student level, the decision of how to best assess an LEP student should be based on several factors, including level of English proficiency, primary language of instruction, level of literacy in the native language, and number of years the student has received academic

instruction in English. The appropriate form of assessment might be assessing the student orally or in writing in his or her native language; providing accommodations such as a bilingual dictionary, extra time, or simplified directions; using an assessment that has been stripped of non-essential language complexity; or administering an English language assessment orally.

Exemptions

Title I does not permit States to exempt any student subgroup from their final assessment systems, though individual exemptions may be permitted by the State in extraordinary circumstances such as medical emergency or parental insistence. If the State exempts any students from the assessments, it must describe the exemption criteria and process. The number of exemptions from the assessment should be minimal and should be based upon reasonable criteria. Furthermore, the State should explain the procedures followed in documenting which students are not assessed, including auditing and record-keeping activities. The State should explain how it plans to reduce the number of exemptions and how it will verify that policies designed to increase student participation in the assessment system produce the intended effects.

In the case of children with disabilities, the final regulations implementing IDEA require that the IEP team have the responsibility and the authority to determine what, if any, individual accommodations or modifications in the administration of state assessments are needed in order for a particular child with a disability to participate in the assessment. Likewise, it is the IEP team that must determine whether a child will not participate in a particular state assessment of student achievement (or part of an assessment) and if not, how that child will be assessed.

Technical Considerations

When assessment procedures are altered, it is critical to ensure that scores, decisions, and judgments based on these assessments are fair, reliable, and valid. The criteria for technical quality outlined in Part II E, "Professional Standards of Technical Quality," apply to modified, accommodated, and alternate assessments. The issue of fairness is one that is particularly salient in the area of assessment accommodations. Accommodations should provide students with the same opportunity to demonstrate their knowledge and skills as students who do not need an accommodation. For example, if language-related accommodations provide students such an opportunity, an LEP student with the same level of knowledge and skills in mathematics as a non-LEP student will achieve the same proficiency level on the assessment.

4. Preparing a State submission of evidence

States should submit three types of documentation as evidence that they have met these requirements. First, they should describe their analysis of the diverse learning needs of the entire student population (including contextual information such as the number of students with disabilities, the number of students needing accommodations, the languages spoken by students in the State, and the number of students in each language category) and the implications for making assessments accessible. Second, they should document policies and strategies that they have implemented to ensure that all students are part of the assessment system. Third, they should submit data on the percentage of students participating in the assessments, explanations for those who were not included, and strategies for including them in the future.

5. Questions for reviewers

Peer reviewer questions	Desirable Evidence	Incomplete or Unacceptable Evidence
B1. Do the State data on assessment participation rates indicate that virtually all students are included in the assessment and that their scores are used to evaluate school and district progress?	Reviewers will look for information describing the State's students with disabilities and limited English proficient students and their rates of participation in the State assessment system, as illustrated in the chart below. Peer reviewers will look for substantial evidence that all students are assessed and their performance is reported. They will look for effective strategies that are being developed and implemented to include any students who have been excluded to date. Evidence also might include descriptions of State policies that provide incentives for including and sanctions for excluding students with disabilities and limited English proficient students from the assessment.	A large number of students with disabilities are excluded from the assessment system, and the State does not have plans for initiating procedures for including these students. Although students with disabilities and limited English proficient students are included in the system, results of their assessments are excluded from measures of school and district progress. The State has adopted policies permitting categorical exemption of students with disabilities and LEP students from the statewide assessment system.

Information to Determine the Need for Test in Language(s) Other Than English

Primary Languages in Grade ___	Number of Limited English Proficient Students
Language 1	
Language 2	
... etc.	

Participation Information for Grade ___

General	Number
Total student population	
Total students with disabilities (IEP & 504)	
Total limited English proficient students	

Participation	Included in Assessment	Included in Measures of Progress
Number of students with disabilities included in State		

assessment **without** appropriate accommodations		
Number of students with disabilities included in State assessment **with** appropriate accommodations		
Number of students with disabilities tested with other State standards-based assessments (beyond accommodations; e.g., alternate assessment)		
Number of limited English proficient students included in State assessment **without** appropriate accommodations		
Number of limited English proficient students included in State assessment **with** appropriate accommodations		
Number of limited English proficient students tested with other State standards-based assessments (beyond accommodations; e.g., alternative, non-parallel test)		

Exemptions and Exclusions	**From Assessment**	**From Measures of Progress**
Number of students with disabilities excluded		
Number of limited English proficient students excluded		

Peer reviewer questions	Desirable Evidence	Incomplete or Unacceptable Evidence
B2 What policies does the State have for including students with disabilities in their assessment system? Does the State policy result in participation rates that provide meaningful data on how well students with disabilities are performing relative to State standards? What policies are provided regarding appropriate accommodations for students with disabilities and the use of alternate assessments?	Peer reviewers will look for substantial evidence that the State has considered the needs of students with disabilities in both the development and implementation phases of its assessment; has effective policies in place for using appropriate accommodations; and has policies to ensure that IEP teams are involved in determining assessment accommodations and whether an alternate assessment is necessary. In addition to counts of participation illustrated in B1, evidence might include State guidelines for appropriate accommodations, handbooks for IEP teams, and other policy documents.	No accommodations are offered for students with disabilities. The State does not have policies for assessing students with disabilities who are excluded from the State assessment(s).
B3. Does the State have a policy in place for maximizing the inclusion of LEP students in the statewide assessment? Does the State policy result in participation rates that provide meaningful data on how well LEP students are performing relative to State standards? What policies are provided regarding appropriate accommodations and linguistically accessible assessments for LEP students?	Reviewers will look for evidence that the State has conducted an analysis of its LEP student population and what their learning needs are, including the use of measures of language proficiency; developed strategies to ensure that they are tested appropriately; and implemented statewide policies or guidelines for appropriate accommodations for LEP students Ideally, peer reviewers would like to see evidence that the State considered the needs of LEP students in the development and design of the State assessment so that it would provide valid and reliable results even with accommodations. If this has not occurred, then strategies to provide appropriate accommodations and multiple measures must be clearly described.	State policies allow the exclusion of LEP students from participating in the State assessment and measures of program progress. The State has not investigated approaches for providing linguistically accessible assessments for LEP students. The State has no procedures for assessing excluded LEP students' achievement in relation to State content and performance standards. The State offers only an assessment in English without accommodations to students who have recently arrived in the U.S. and are not proficient in English.

Peer reviewer questions	Desirable Evidence	Incomplete or Unacceptable Evidence
B4 Does the State offer native language assessments for some LEP populations? Are policies in place to ensure that they are used appropriately? If not, why not? Is it practicable to offer these in the future? Does the State require that staff conducting native language assessment possess adequate proficiency in the native language? Are they adequately prepared and trained in the assessment procedure?	Peer reviewers will look for evidence that the State is at least making a Spanish language version of the test available (since over 70% of LEP students are Spanish speakers), unless the State has a very small Spanish-speaking population. Reviewers will look for evidence that the State has additional strategies for adopting or developing native language assessments where appropriate.	The State does not offer native language assessments and has not shown that provision of linguistically accessible assessments to LEP students would be impracticable.
B5 Do accommodations offered to students with disabilities and LEP students reflect the instructional approaches used with those students?	Reviewers will look for evidence of standard procedures and guidelines for determining which accommodations are appropriate for individual students. Evidence should show how accommodations reflect the ways students learn content.	Students are provided accommodations that are unrelated to instructional approaches routinely used in their instruction (e.g., an audiotaped assessment administration when a student routinely reads print material).
B6 Do the accommodations offered to students with disabilities and LEP students provide a means for making valid inferences about the knowledge and skills of these students? Has the State investigated the technical quality of the accommodated scores?	Reviewers will look for evidence that appropriate accommodations have been selected or developed in such a manner that valid inferences can be made about student proficiency in relation to State standards. Such evidence might be found in descriptions of expert review of and recommendations for appropriate accommodations and studies conducted on the effects of appropriate accommodations on student scores.	The State has no rationale, either judgmental or empirical, for the accommodations it offers and the accommodations it prohibits.

Peer reviewer questions	Desirable Evidence	Incomplete or Unacceptable Evidence
B7 Does the State monitor the application of inclusion policies at the local level?	Peer reviewers will look for evidence that the State has a means of ensuring that inclusion and accommodation policies are applied consistently and appropriately across the State. Evidence might include descriptions of training in using inclusion and accommodation guidelines, a description of monitoring procedures, or a report of the results of monitoring application of State guidelines.	The State does not provide information on how to apply its guidelines and does not monitor how closely LEAs follow the guidelines.

Examples of Sources of Evidence for Part I

Types of evidence States might use to document progress pertaining to General Characteristics of the Assessment System are described below. This list does not include all possible types of evidence; rather, it is designed to serve as a source of ideas for States as they prepare their evidence for review.

Sources of Evidence	Content, grade levels, & administration	Inclusion
Description of assessment system, including subject areas, grades assessed, and frequency of administration	X	X
Sample score reports	X	X
Assessment development materials (e.g., assessment blueprints, item specifications)	X	X
Criteria/models for LEA adoption/development of assessments	X	X
Description of assessments or tracking procedures for schools without grades covered by the State assessment	X	X
Administration manual that includes descriptions of allowable accommodations for students with disabilities and guidelines for the inclusion or exemption of limited English proficient students	X	X
Description of procedures for determining whether students with disabilities (both IEP and 504) and limited English proficient students are provided with appropriate accommodations/translations or exempted from the general assessment		X
Description of development of alternate assessment procedures for students with disabilities		X
Description of development of linguistically accessible assessments		X
Description of instruments used to assess language proficiency of limited English proficient students		X
Documentation of number of exemptions from State assessments		X

PART II – THE CORE OF THE ASSESSMENT SYSTEM

PART II – C: Assessments Must be Aligned to Standards

1. Requirement-Legal Citation

> The State assessment shall –
> Be aligned with the State's challenging content and student performance standards and provide coherent information about student attainment of such standards.
> (Sec. 1111(b)(3)(B))

2. Intent and purpose

The intent of this requirement is to ensure that assessments reflect what students are expected to know and be able to do. Such assessments will help guide educators in measuring student progress and making necessary alterations in their teaching and learning strategies to help all students master challenging State standards. For the purposes of this review, alignment is defined as the degree to which assessments provide valid and accurate information about the performance of all students in an academic content area at the desired level of detail on the State's content standards.

3. Abbreviated Description

Demonstrating that an assessment system is aligned to State content and student performance standards requires more than simply determining whether all the items on the assessment can be matched to one or more standards; the converse must also be probed. In other words, States should also determine whether the State assessment adequately measures the State's standards. This can be accomplished by analyzing how well the State assessment measures State standards along the following dimensions:

- **Comprehensiveness:** Does the assessment reflect the full range of the standards? If not, does it sample enough to make relevant inferences about student performance on the entire set of standards? Is it complemented by other measures, such as another test or local measures that provide information to educators on the other standards?
- **Emphasis:** Does the assessment reflect the same degree of emphasis on the different content standards as reflected in the standards documents?
- **Depth:** Does the assessment reflect the cognitive depth of the standards? In other words, is the assessment as cognitively demanding as the standards?
- **Match with performance standards:** Does the assessment provide scores that reflect the meaning of the different performance standards?
- **Clarity for users:** Is the alignment between the standards and the assessment clear to all members of the school community?

4. Full Description

Alignment has many meanings in education. In one sense, it is the core idea underlying standards-based school reform; reform cannot happen unless all parts of the system come together--not just standards and assessments but virtually all the other components of an educational system including teaching strategies, instructional materials, and professional development.

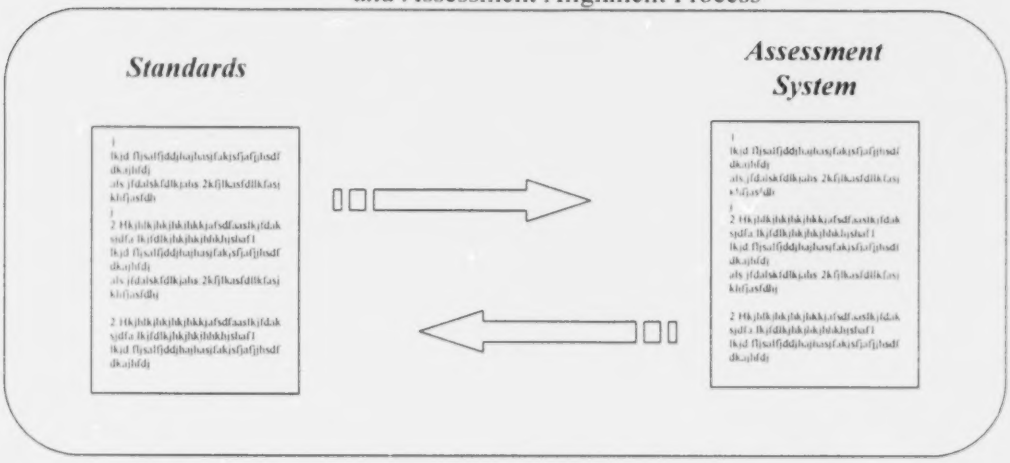
Focusing on its systemic aspects, Webb defines alignment as the degree to which expectations and assessments are in agreement and serve in conjunction with one another to guide the system in ensuring that students learn what they are expected to know and do (Webb, 1997). More specific to the operational review process outlined in this document, we will define alignment as the degree to which assessments report valid and accurate information about the performance of all students in an academic content area at the desired level of detail on the State's content standards.

Each State must present evidence that their assessment system is aligned to their standards. This general statement means many specific things, and it means different things for different States, depending on the design of their State systems. It means something different to States that custom-developed their assessments to match the standards, in contrast to those States that adopted an assessment based on an alignment study. However, in both cases it includes the dual and sometimes overlapping processes of obtaining alignment, as well as verifying it. In fact, in some cases it includes the process of re-verification, if changes in tests were made to improve alignment.

Since this document is for peer reviewers, it only provides glimpses of procedures that States can use to achieve or verify alignment. Fortunately, one of the CCSSO's State Collaboratives on Assessment and Student Standards (SCASS) has developed a document that describes and illustrates several approaches to alignment and alignment verification (LaMarca, Redfield and Winter 1999). A CCSSO project led by Blank and Webb also developed rating procedures for examining alignment, and applied them to the standards and assessments from four States (Blank and Webb, 1999).

It seems obvious that alignment is a two-way process, especially for States that choose to select an existing assessment. It is not sufficient that a State determines that all the items on the assessment can be matched to one or more standards; the converse must also be probed, "Are all the standards adequately assessed?" The following visual may help to illustrate this basic point.

Figure 1. The Two-Way Nature of the Standards and Assessment Alignment Process



Another way to make this point is with a Venn diagram. Figure 2 illustrates, hypothetically, what might have happened when a given State compared its standards to a given assessment.

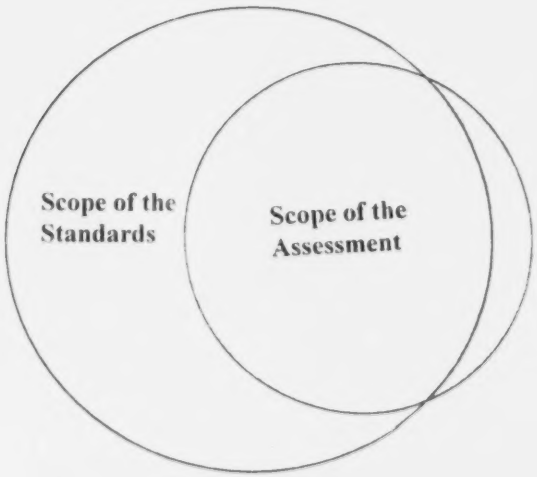


Figure 2. A Visual Display of Partial Alignment Between State Standards and State Assessment

In the Venn diagram above, the assessment almost completely matches the standards, that is, only a small portion of the assessment[1] relates to knowledge and skills that are not referred to in

[1] The shaded circle is meant to represent all aspects of the assessment, for example, both a norm-referenced test and a State-developed writing assessment. It might also include certain local assessments, pursuant to the following discussion.

BEST COPY AVAILABLE

the standards. But, going the other direction, the story is very different--about half of the standards are not assessed.

Facets of Alignment

To satisfy the definition given above, a State's assessment system must

a) reflect the full range (**comprehensiveness**) of the standards;
b) reflect the relative **emphases** on the different content standards;
c) reflect the **(cognitive) depth** of the standards;
d) provide scores that reflect the meaning of the different **performance standards**; and
e) make it **clear and transparent** to all members of the school community how the standards and assessments are aligned.

The following paragraphs further describe these five elements.

Content Alignment--Comprehensiveness.

Comprehensiveness implies that all standards are to be assessed. This idea of comprehensiveness is addressed in Webb's framework as the criteria of "categorical concurrence" and "range of knowledge correspondence." It means that standards and assessments cover a comparable span of topics and ideas within categories, and do so at the specified level of detail (Webb, 1997). In the face of challenging content standards, it is unlikely that a single assessment instrument will provide the simultaneous breadth and depth necessary for a fully aligned system.

The law requires only that the assessments

- be aligned with the State's challenging content and student performance standards;
- include the same knowledge, skills, and levels of performance expected of all children; and
- measure performance in at least mathematics and reading/language arts.

This leaves a great deal of flexibility to the States. A State must decide whether all standards will be assessed at all grade levels or whether to assess at selected benchmark grades; whether all standards will be covered in a single assessment administered at one point in time; whether all standards will be assessed in equal depth--some standards may be more complex or more important than others; whether all standards will be addressed by the State assessment or whether to leave the assessment of selected standards to their LEAs (either all the standards for some content areas, or for certain strands within the content areas of reading/language arts and mathematics).

Determining how all of a State's standards should be assessed depends upon the structure of the standards. Some States have broad standards, fewer than 10 per content area in any grade level. Other States have more detailed standards, in some cases over 30 per content area. In most cases, the process of translating the standards into assessment blueprints will lead naturally to decisions about the relative depth and breadth of sampling of each standard. For a State with broadly defined content standards, the assessment will probably require several items matched to each standard. For a State with detailed content standards, sampling within chunks of content may provide data that can be used to make inferences at the desired level.

States may choose to--

- assess certain standards only at certain grade levels, provided that the State has a credible rationale for doing so;
- assess certain content areas at the State level and others at the local level, provided that the identification of schools for Title I program improvement is based, at a minimum, on the content areas of reading/language arts and mathematics;
- assess selected standards within the content areas of reading/language arts and mathematics as part of its State assessment, and allocate the remaining standards within reading/language arts and math to the LEAs for assessment. If the components or content strands that are assessed at the local level are included in the State's definition of adequate yearly progress, the State must monitor the local assessments to assure objectivity, accuracy, and comparability.

When documenting the comprehensive aspects of alignment between standards and the State assessment system, the State should describe--

- the relationships between the structure of the standards and the structure of the assessments;
- the rationale for the overall alignment strategy, including a rationale for any standards either not assessed or not reported as part of the State assessment; and
- the manner in which each standard is assessed, whether at the State, district, school, or classroom level
  - the type of information the State collects pertaining to each standard, and
  - how the State monitors the quality of the assessment data collected at the local level, for all assessments that are part of the statewide Title I system.

Appendix B describes the decision-making process required to develop an assessment system that addresses all standards. It includes consideration of the purposes and uses of the assessment, the jurisdictional level at which the assessment is conducted, and the relative appropriateness of assessing a given standard in a formal manner on a statewide basis.

b. Content Alignment--Emphasis.

An aligned assessment will cover the knowledge and skills specified by the content standards with the same degree of emphasis as specified or implied by the standards. This is essentially a matter of weighting, a matter of making sure that standards that are judged more important than others get more weight in the computation of an overall score (whether at the school level or the student level). The most straightforward indicator of emphasis is the number of test questions per standard or subset of standards. It is important that the relative emphases be obvious to teachers if the assessment is going to support the aims of the standards.

This use of number or proportion of items per standard is related to the use of different types of assessment formats, partially because different types of assessment exercises take different amounts of time. Performance items typically take more time, but also yield more information (both in a pedagogical sense and a statistical sense). The amount of time devoted to assessing different standards can also signal relative importance. States will need to work out a balance of these different factors.

c. Content Alignment--Depth

It may seem that if alignment is alignment, a true match of standards and assessments would automatically ensure a match on other criteria such as emphasis and depth. In the real and slightly messy world of alignment, however, it is important to verify that the assessments reflect the degree of cognitive complexity and level of difficulty of the concepts and processes described in the standards. Webb puts it this way: "*...what is elicited from the students on the assessments is as demanding cognitively as what students are expected to know and do as stated in the standards*" (Webb, 1999, p.7). The meaning of "cognitively demanding" is broad, including how well students should be able to transfer their knowledge to different contexts and how much prerequisite knowledge they must have in order to grasp more sophisticated ideas. Moreover, the law calls for the assessment of complex skills and understanding.

LaMarca, Redfield, and Winter describe feasible strategies to document the alignment of tests to content standards. At the very least, a State can study the nature of the verbs used in the standards and look for their manifestations in the assessment. A State might begin by categorizing the complexity and set of cognitive demands specified or implied by each standard, then develop a set of criteria for review.

d. Alignment to Performance Standards.

An aligned assessment reflects the nature of the student performance described in the performance standards, as well as the content standards. The Council of Chief State School Officers and the US Department of Education recently produced a handbook on performance standards. *Handbook for the development of performance standards: Meeting the requirements of Title I* (Hansche, 1998) that sets forth the essential characteristics of performance standards as a key component of a standards-based assessment system. Performance standards describe the level(s) of acceptable performance, specify those levels in operational assessment terms, and provide a mechanism for reporting the results in terms of the proportion of students who meet the standards. Key elements include--

- performance descriptors--narrative descriptions of performance at each level; and
- exemplars--examples of student work from a representative sample of all students that illustrate the full range of performance at a level.

The implications are obvious: the content of the assessment must match the knowledge and skills described in the performance descriptors for each performance level. Furthermore, any tasks and student work used to illustrate the meaning of the descriptors must reflect the actual tasks used in the assessment.

e. Clarity and Transparency of the Alignment.

The alignment between standards and assessments needs to be reflected in various documents available to teachers, students, and parents. These users ought to be able to see easily how the meaning and the relative weight of the different standards are reflected in the assessments. Assessment reports can help to communicate this alignment, but it is likely that other documents will be needed also.

Conducting the Alignment Process

Webb (1999) suggests that both content experts and people knowledgeable about a State's standards and assessments serve on review panels as part of the alignment and alignment verification process. These reviewers need training in the review process and should be monitored periodically throughout the process to ensure that they are applying the review criteria appropriately.

5. Preparing a State submission of evidence

States will be expected to describe how they are addressing each of the five aspects of alignment described above. The State should provide evidence that it has studied the alignment of the assessment and standards and, if gaps exist, that it has identified additional measures to adequately assess the standards. In some cases, the State may need to focus more on its plans than its progress in addressing these facets of alignment. Peer reviewers would then consider these plans as well as the documentation of what the State has already accomplished. There is no single best way of accomplishing the alignment or documenting the process, but it is reasonable to expect that a broad variety of stakeholders will be involved in the process, and that the assessment blueprints or specifications play a key role.

In States that develop their assessments to fit the standards, the reviewers might expect an independent post hoc review to confirm successful alignment. Other assessment development strategies might call for two alignment reviews done at different times. The first would identify the relative alignment of different ready-made assessment packages, and the second would confirm the process after the problem of any serious gaps in assessing the standards had been addressed. The bottom line is that the responsibility for alignment rests with the State, regardless of the State-local configuration or of the assessment development strategy selected.

6. Questions for reviewers:

Peer reviewer questions	Desirable Evidence	Incomplete or Unacceptable Evidence
C1. What is the State's approach to ensuring alignment of its standards and assessment? What kinds of alignment studies have been done? Who was involved? What methodology was used? What were the findings?	Reviewers will look for a description of the State's approach to ensuring alignment. They will evaluate whether the approach is reasonable and thoughtful. They will be looking for evidence that the State is taking a coherent approach to ensuring that its tests reflect what the State has determined students need to know and do. This almost surely will involve some type of alignment study.	A checklist showing that all of the assessment items match one or more standards A study that did not involve content experts, that examined the alignment only at a very global level, or that failed to ensure objectivity in the process
C2. How is the State ensuring that its assessment system reflects its content and performance standards in terms of	Reviewers will look for evidence from the assessment plan, the assessment blueprints and/or item/task specifications that the State considered how all content standards would be assessed or how domain sampling	An assertion of comprehensiveness without documentation matching both assessments to standards and standards to assessments.

Peer reviewer questions	Desirable Evidence	Incomplete or Unacceptable Evidence
comprehensiveness and **emphasis**?	would lead to valid inferences about student performance on the standards. They will look for descriptions and evidence that (a) the full scope of the standards and their differential emphases are reflected in the blueprints and that (b) the assessments match the blueprints. They will expect to see that impartial experts were involved in the process.	
C3. How is the State ensuring that its assessment reflects its content and performance standards in terms of **depth** and **match with performance standards**? How is the State ensuring that its assessment covers the range of cognitive complexity of its standards, not just the basic skills? How is the State ensuring that the assessments actually reflect the types of student performance called for in performance standards?	Reviewers will look for a description and evidence that cognitively complex standards are adequately assessed. As in comprehensiveness and emphasis, reviewers will look for evidence that the blueprints reflect the standards that call for higher order or cognitively complex skills, and that the assessments match the blueprints.	--Evidence that some assessment items measure higher order thinking, but not showing that most of the standards that call for higher order thinking are adequately assessed --Using a methodology that does not examine whether the more complex standards are assessed or whether the assessment tasks parallel the illustrative tasks in the performance standards
C4. How clearly has the State identified any gaps or weaknesses and what is it doing to improve the alignment of its assessment and standards?	A discussion of the gaps found and a description of the strategies that the State is putting into place to address them such as: • adding items to the assessment • adding multiple measures • adding a writing test • adopting the longer version of a test	--Conducting alignment studies, even high quality studies, but not describing steps taken or planned to strengthen the alignment if gaps were found
C5. If the State system consists of several assessments or draws upon assessment data from several sources, is there a coherent design that shows how all the standards are assessed?	Reviewers will look for descriptions of the State's assessment system plan which describes the ways in which different assessments provide for alignment, and • how the results from the different assessments are reported, separately or combined (if and when that is appropriate); • how the results from the different	--Simply listing the different assessments without showing how they fit together to form an assessment system --Indicating that some of the standards are assigned to the schools for assessment using their own instruments, without showing how this process leads

Peer reviewer questions	Desirable Evidence	Incomplete or Unacceptable Evidence
	assessments are to be interpreted by the users; • how comparability issues are handled (even though this is mainly dealt with under "technical quality"); and • the different roles of local and State personnel in selecting and scoring the assessments, and in interpreting and using the information.	to valid inferences about the effectiveness of programs in schools across the State
C6. How is the alignment of the assessment and the standards communicated? Is it clear to educators and parents what is being assessed and how it relates to the standards?	Reviewers will look for ways the State has used various documents such as manuals, bulletins, reports of results, and website displays to show the alignment and communicate this information both to educators and the public.	--Indicating or implying that there really is no easy way for teachers or the public to see how or how well the assessments match the standards

Appendix C provides additional illustrations of types and sources of evidence that a State might consider when studying alignment. It is also a good source for the types of evidence that peer reviewers might look for within each category of alignment.

PART II – D: Professional Standards of Technical Quality

1. Requirement--Legal Citation

> State assessments shall –
> Be used for purposes for which such assessments are valid and reliable, and be consistent with relevant, nationally recognized professional and technical standards for such assessments. (Section 1111(b)(3)(C))

2. Intent and purpose

The intent of this requirement is to ensure that the assessment data that are used to hold schools accountable are indeed technically sound and meaningful. Ensuring technical quality of assessments is an ongoing task that will continue as long as assessments are in place. However, for the purposes of this review, each State needs to document that its assessments are technically adequate and that it has taken reasonable steps to ensure that results are used in a manner that is technically sound.

3. Abbreviated Description

Although the law mentions only the two most well known technical characteristics, validity and reliability; a number of additional requirements are considered essential. Other criteria discussed in this section include fairness/equity; comparability; administration, scoring, analysis and reporting processes; and interpretation and use.[2] Peer reviewers will look for evidence of technical quality along six dimensions.

a. Validity – "the appropriateness, meaningfulness, and usefulness of the specific inferences made from test scores" (Standards, 1985, p. 9). Peer reviewers will look for evidence that:

- the State has considered whether the inferences drawn from the assessment are appropriate and meaningful;
- the State has examined construct validity (whether the assessment actually measures the content and performance standards in question); and
- the State has examined consequential validity (the validity as judged by the long-term impact of the results).

b. Reliability – the level of consistency, stability, and accuracy of the assessment. The *Standards* explains reliability as follows: "Fundamental to the proper evaluation of a test are the identification of major sources of measurement error, the size of the errors resulting from these sources, the indication of the degree of reliability to be expected between pairs of scores under

---

[2] We might have listed self-examination and continuous improvement as criteria, since they are essential for any system, especially one that exists solely to improve other systems. Although not a formal requirement, States need to take a proactive stance and systematically seek evidence that the system is providing the best possible information in the best way possible.

particular circumstances and the generalizability of results across items, forms, raters, administrations, and other measurement facets" (1985, p. 19.). This is a tall order, a task that is beyond the present practice of many programs. The focus of the reviewers, therefore, will include the adequacy of the plans for, and initial steps taken to, carry out this process.

c. Fairness/accessibility – ensuring that all students have an equal opportunity to show what they can do, in spite of the fact they have different backgrounds, different and complex patterns of abilities that interact with the assessment process itself, and different opportunities to meet the standards.

d. Comparability of results -- from year to year, school to school, and student to student. Given the demands placed on Title I assessments to detect change, especially from year to year, it becomes necessary to consider comparability in designing and developing the assessment, and then in gathering confirmatory data during the implementation phase. Although difficult to implement and to document, States have an obligation to show they have made a reasonable effort to attain comparability, especially where locally selected assessments are part of the system.

e. Administration, scoring, analysis and reporting procedures. Most states take great pains to ensure that the assessments are properly administered, that directions are followed, that test security requirements are clearly specified and followed, and that all students are assessed. Nevertheless, it is important they document the ways in which they ensure that their system does not omit any of these basics.

f. Interpretation and use – ensuring that users of the assessment data have the support needed to draw the most appropriate interpretations and use the results in the most valid ways.

3. Full description

Only the most commonly agreed-upon principles and criteria related to technical quality are presented here.[3] Most of these are discussed in greater detail in two authoritative documents in the field, the *Standards for Educational and Psychological Testing* (1985) and *Educational Measurement* (Linn, 1989). Reference is also made to the draft of the next revision of the *Standards for Educational and Psychological Testing*[4] (in press) which is scheduled for publication as soon as the three sponsoring organizations[5] give their approval, which is expected later this year.

---

[3] The reader may notice that the various criteria, especially validity and reliability, refer at times to individual assessment issues and at times to issues related to the use of group-level summaries. The focus of the peer reviewers is on both individual and group issues, depending on the particular purpose and use of the assessment information in question..

[4] As this document was being prepared, the authors consulted draft versions of the revised *Standards for Educational and Psychological Testing* in an effort to assure consistency with the 1999 Standards.

[5] The American Educational Research Association, The National Council on Measurement in Education, and the American Psychological Association.

a. Validity

This complex topic is often simplified in textbooks in the form of the quasi-tautological question, "Does the test measure what it purports to measure?" This turns out to be a difficult question to answer, sometimes leading to considerable controversy. This discussion of validity recognizes three relatively recent major conclusions about the definition of this elusive concept.

- The focus of validity is not really on the test itself, but on the **inferences** drawn from the results that it yields.
- All validity is really a form of "**construct** validity."
- In validating an assessment, one must also consider the **consequences** of its interpretation and use.

Drawing Inferences.

Over the years the focus has been on different types of validity, such as content validity or concurrent validity. It is now agreed, however, that validity is a global concept centering on the inferences that are drawn from a set of findings by a given user in the light of the purpose of the assessment. The *Standards for Educational and Psychological Testing* underscores this definition:

> Validity refers to the appropriateness, meaningfulness, and usefulness of the specific inferences made from test scores. Test validation is the process of accumulating evidence to support such inferences.....Although evidence may be accumulated in many ways, validity always refers to the degree to which that evidence supports the inferences that are made from the scores. The inferences regarding specific uses of a test are validated, not the test itself. (1985, p. 9)

The draft revision of the *Standards for Educational and Psychological Testing* only serves to underscore this emphasis on the drawing of inferences. It goes on to assert that the various types of validity are really types of evidence that can be used to confirm the appropriateness of drawing certain types of inferences about student performance on the basis of test scores. It recasts the traditional types of validity in terms of types and sources of evidence, all of which pertain to construct validity. It speaks of four broad categories of evidence: (1) evidence based on the assessment's relation to other variables, (2) evidence based on student response processes, (3) evidence based on test content, and (4) evidence from internal structure.

Construct Validity

The second major transformation is a natural sequel to the first. That is the realization that "construct validity" is not just one of many types of validity--it *is* validity. All validity evidence and arguments are focused on the basic question, "Is the assessment tapping the concept, skill or trait in question? Is it really measuring mathematical reasoning or reading comprehension? A variety of types of evidence and analyses can be used to answer such a question--none of which provide a simple yes/no answer.

The reader is reminded of the Venn diagram used in the alignment section. It illustrated the omission of some aspects of content by the assessment, and the undesirable inclusion of other

aspects of learning that were outside the scope of the standards. In the parlance of construct validity, the sections not assessed that should be are known as "construct under-representation" and the topics that are assessed, perhaps inadvertently, are known as "construct irrelevant variance." At the level of a content match, as in alignment, it is relatively easy to identify both types of mismatch and their magnitude.

Determining whether an assessment is actually measuring what the State intended is a more difficult matter. The draft *Standards for Educational and Psychological Testing* illustrates how various threats to construct validity might undermine the meaning of scores on a reading comprehension test. If a student has a strong emotional reaction to the reading passage, for example, it is easy to see how the results might not be a valid estimate of his/her reading ability. Similarly, if the assessment calls for students to write a long response to explain their answers, the results for some students might be distorted by their writing ability.

Distinguishing what is measured from what is not measured often involves the use of triangulation--a process of reasoning from diverse sources of evidence, including the four mentioned below.

1) Using evidence based on test content (content validity). It is now widely recognized that content validity is one facet of construct validity that appears mainly in the validation of achievement tests. In fact, the question is often posed, "Is construct validity really separate from content validity?" Messick (1988) answers negatively:

> Typically, content-related inferences are inseparable from construct-related inferences. What is judged to be relevant and representative of the domain is not the surface content of test scores but the knowledge, skill, or other pertinent attributes measured by the items or tasks (1988, p. 38).

Content validity, that is, alignment of the standards and the assessment, is important but not sufficient. States must document not only the surface aspects of validity illustrated by a good content match, but also the more substantive aspects of validity that clarify the "real" meaning of a score.

2) Using evidence of the assessment's relationship with other variables. One approach is to document the validity of an assessment by confirming its positive relationship with other assessments or evidence that are known or assumed to be valid. For example, if students who do well on the assessment in question also do well on some trusted assessment or rating, such as teachers' judgments, it might be said to be valid.

It is also useful to gather evidence about what a test does *not* measure. The *Standards for Educational and Psychological Testing* propose that:

> When a test is proposed as a measure of a construct, evidence should be presented to show that the score is more closely related to that construct when it is measured by different methods than it is to substantially different constructs (1985, p. 15).

This means, for example, that a test of mathematical reasoning should be more highly correlated with another math test, or perhaps with grades in math, than with a test of scientific reasoning or a reading comprehension test. The most common--and complicated--example is found in this very area, teachers are frequently concerned that tests of mathematical reasoning might actually measure reading comprehension since students must be able to understand the problem, which is usually presented in narrative form. Although students obviously need to be able to read well to understand the math task, the validation challenge is to marshal evidence that students who do well on the assessment are not relying on their reading ability to answer the questions, and simultaneously--if possible--to confirm that students who are less skilled readers are not hindered in demonstrating their mathematical understanding.

3) Using evidence based on student response processes. The best opportunity for detecting and eliminating sources of test invalidity occurs during the test development process. Items obviously need to be reviewed for ambiguity, irrelevant clues, and inaccuracy. More direct evidence bearing on the meaning of the scores can be gathered during the development process by asking students to "think-aloud" and describe the processes they "think" they are using as they struggle with the task. Many states now use this "assessment lab" approach to validating and refining assessment items and tasks.

4) Using evidence based on internal structure. A variety of statistical techniques have been developed to study the structure of a test. These are used to study both the validity and the reliability of an assessment. The well-known technique of item analysis used during test development is actually a measure of how well a given item correlates with the other items on the test. If an item gets a high index, we say it is a good item, meaning that students who get it right also tend to do very well on most of the other items. This practice actually helps ensure a focus to the assessment. It means that although a reading comprehension test consists of items that measure different aspects of comprehension, there is a core focus that helps ensure the reliability of the assessment. Newer technologies including generalizability analyses are variations on the theme of item similarity and homogeneity.

Other techniques are used to show whether there are certain clusters of items. Whether, for example, the items measuring mathematics computation tend to "hang together" and the items in concepts and problem solving tend to form a relatively separate cluster. Although the number of clusters that these statistical methods are able to identify is nearly always fewer than the number of content categories used in test development, it can still be a useful exercise as part of the package of construct validation techniques. A combination of several of these statistical techniques can help to ensure a balanced assessment, avoiding on the one hand, the assessment of a narrow range of knowledge and skills but one that shows very high reliability, and on the other hand, the assessment of a very wide range of content and skills, triggering a decrease in the consistency of the results.

Multiple measures. One purpose of multiple measures is to ensure validity in both the relatively superficial sense of content validity and the deeper aspects of construct validity. Different types of measures and tasks, including the use of different testing formats, are needed to assess different content standards and to measure the different types of knowledge

and skill represented by those standards. Multiple measures can also play a role in ensuring the validity of interpretations of performance for diverse populations.

Consequential Aspects of Validity.
The third major shift in recent thinking is that the evaluation of an assessment must also look at the consequences of the assessment, including the application of the results. Messick (1989) points out that test interpretation and use are different functions, and that the impact of an assessment can be traced either to an interpretation or to how it is used. He also notes that if we are trying to see if an assessment is "doing the job" it is quite natural that we look at the consequences of the assessment. In fact it is rather amazing, in retrospect, that this is a new realization!

> The point is that the functional worth of the testing depends....on the consequences of the outcomes produced, because the values captured in the outcomes are at least as important as the values unleashed in the goals (Messick, 1989, p. 85).

Furthermore, as in all evaluative endeavors, we must attend not only to the effects, but to the side effects.

> Judging validity in terms of whether a test does the job it is employed to do...requires evaluation of the intended and unintended social consequences of test interpretation and use (Messick, 1989, p. 84).

The array of possible consequences for individual students or groups of students is wide. The analysis of consequences is often focused on the unintended or unnoticed consequences of the assessment. The disproportional placement of certain categories of students in special education is an example of an unintended--and negative--consequence of what had been considered proper use of instruments that were considered valid. More recently, assessment has been used as a policy tool to help focus instruction on certain valued outcomes. On the other hand, if the assessment narrowly focuses on certain types of skills, it can have a negative impact on instruction--and learning. Messick (1989) chose to focus on this very example of unintended consequences:

> [Consequential aspects of validity]....require evaluation of the intended and unintended social consequences of test interpretation and use. For example, the use in educational achievement tests of structured response formats such as multiple-choice (as opposed to constructed responses) might lead to increased emphasis on memory and analysis in teaching and learning at the expense of divergent production and synthesis (1989 p. 39).

b. Reliability

The term "reliability" is usually defined with synonyms such as consistency, stability, and accuracy. These terms all relate to the problem of uncertainty in making an inference about a score. As reflected in the *Standards for Educational and Psychological Testing*, the field now

treats reliability as a study of the many sources of unwanted variation in assessment results.[6] Those responsible for developing and operating State assessment systems are obliged to (1) make a reasonable effort to determine the types of error that may (unwittingly) distort interpretations of the findings, (2) estimate their magnitude, and (3) make every possible effort to alert the users to this lack of certainty. The *Standards for Educational and Psychological Testing* puts it this way:

> Fundamental to the proper evaluation of a test are the identification of major sources of measurement error, the size of the errors resulting from these sources, the indication of the degree of reliability to be expected between pairs of scores under particular circumstances and the generalizability of results across items, forms, raters, administrations, and other measurement facets (1985, p. 19).

This is a tall order, a task that is beyond the present practice of many programs. The focus of the reviewers, therefore, will include the adequacy of the plans for and initial steps taken to carry out this process.

The reliability of an assessment, or lack of undesirable variability, is a function of many factors. Three of the factors most relevant to State assessment are briefly discussed below.

Sampling. Assessment is essentially a sampling problem. That is, it is matter of sampling from a domain of all the skills that could be assessed; therefore, students need the opportunity to take a sufficiently large sample of items or tasks in order to yield a stable estimate of their level of performance. The relationships between number of items each student takes and the consistency of the scores are well known; for example, the amount of improvement in reliability is great when moving from a small to a medium number of items, but after a certain number of items, the improvement is relatively trivial. As mentioned under validity, the law calls for the use of multiple measures. The implications of multiple measures for reliability are obvious; increases in score consistency and stability result from the administration of additional exercises, whether they are administered at one time or over a period of time (yielding other useful information in the process). Fortunately, when using assessment results for school accountability the problem is substantially reduced, since errors in estimating individual student performance tend to cancel out at the group level. The use of matrix-sampling increases the stability of the results at the school level still further.

Level of challenge. In order to show what they can do, students need to respond to tasks that are within their range of knowledge and their skill level. If the assessment taps content that students have not been exposed to, they will not respond or will respond randomly. Similarly if the level of the assessment is far below their level of functioning, their scores will be less accurate, either over- or under-estimating their actual performance.

Rater accuracy. A third issue has come to the forefront in recent years with the increasing use of essay tests and other performance assessments: the degree of agreement of those rating the

[6] And the magnitude of these errors is often larger than has commonly been reported.

results. Even back in 1985, the *Standards for Educational and Psychological Testing* stressed the obligation to report the degree of consistency among the raters[7]:

> Where judgmental processes enter into the scoring of a test, evidence on the degree of agreement between independent scorings should be provided (1985, p. 22).

Reporting level of accuracy.
The information or evidence provided by States on the stability of their assessments will indicate how well these and other issues have been addressed. The traditional methods of portraying the consistency of test results, including reliability coefficients and standard errors of measurement, should be augmented by, if not replaced with, techniques that more accurately and visibly portray the actual level of accuracy (Rogosa, 1995, Young and Yoon, 1999). Most of these methods focus on error in terms of the probability that a student with a given score, or pattern of scores, is properly classified at a given performance level, such as "proficient." For school-level or district-level results, the report would indicate the estimated amount of error associated with the percent of students classified at each performance level. For example, if a school reported that 47% of its students were proficient, the report might say that the reader could be confident at the 95% level that the school's true percent of students at the proficient level is between 33% and 61%. Furthermore, since the focus on results in a Title I context is on growth, the report should also indicate the accuracy of the year-to-year changes in scores.

For reliability, the obligation of the States is two-fold. First, they need to document the reliability of the scores at the student level (unless a matrix-sampling design is employed) and the school level. Second, they need to show that they are taking all reasonable steps to inform, in the most meaningful way possible, the consumers of the student and school reports of the level of accuracy of the results.

c. Fairness/Accessibility

Fairness could well be considered a facet of validity, since it poses the question, "Is the inference that one would draw about a student's performance on this assessment valid, or is there something about the assessment or its interpretation which prevents a clear affirmative answer?" However, fairness is treated separately in order to help ensure that States do not overlook any known trouble spots, and to help them develop an effective plan to identify and eliminate them. It is also treated separately in the draft of the revision of the *Standards for Educational and Psychological Testing*.

Like validity, fairness has been the subject of considerable research and much has been written about it. Nevertheless, it suffers from the lack of a single specific definition. For the purposes of this review document, fairness means that all students have an equal opportunity to show what they can do, in spite of the fact that they have different backgrounds, different

[7] The real issue, of course, is not the "scoring reliability" but rather the overall "score reliability," which includes various other types of error variance as well as scorer reliability.

and complex patterns of abilities that interact with the assessment process itself, and different opportunities to meet the standards.

The draft version of the *Standards for Educational and Psychological Testing* identifies several types or sources of unfairness:

- bias or unequal treatment of students in the assessment process or in the processes of reporting, interpretation or use;
- the lack of opportunity to learn to the standards.

It is especially important that States take steps to ensure fairness for the populations that may have been victims of unfair assessment in the past. These populations include the very target of Title I programs--students from poverty--as well as English language learners and students with disabilities. Many of the most critical issues involved in ensuring fairness for these latter groups were treated in Part I under the topic of assessing all students. The strategies for assessing these students, including assessing students in their primary language and using accommodations and alternate assessments, are still being developed, studied, and refined. Admittedly, there is controversy about their use. The use of any of these strategies at this point does not produce incontrovertible evidence of fairness, validity, reliability, or comparability. Nevertheless, the State must describe the steps it is taking, and its plans for making the assessments as fair as possible. (And the solutions will only come as States try different approaches and provide detailed information on the results of their efforts.)

Unfairness most often appears at four points in the assessment process. These four points might serve as a framework for States to use in attacking the problem--and for reviewers to use in judging the adequacy of their efforts.

- The items or tasks do not provide an equal opportunity for all students to fully demonstrate their knowledge and skills.

  This issue can be addressed through an aligned assessment that provides all students in the system the opportunity to demonstrate their proficiency relative to the content standards. It allows students who have learned the content in different ways, students with disabilities, and students who are English language learners to fully demonstrate their knowledge and skills. Assessments should allow for
  --different ways of expressing competency and responding to tasks, (i.e., accessibility Note: Appendix D provides additional details on the important area of accessibility)
  --the use of accommodations and modifications,
  --the screening for bias and irrelevant factors, and
  --the empirical study of items and tasks.

- The assessments are not administered in ways that ensure fairness.
- The results are not reported in ways that ensure fairness.
- The results are not interpreted or used in ways that lead to equal treatment.

(Note: These four points are illustrated further in Appendix D, Summary of Technical Quality Criteria and Illustrative Evidence)

Finally, States are reminded of the requirement for the use of multiple measures, which can be a part of the total solution. Students that may not be able to demonstrate their skills effectively on one type of assessment may do very well on another.

d. Comparability of Results (not fiscal comparability!)[8]

Many uses of State assessment results assume comparability of different types: comparability from year to year, from student to student, and from school to school. To some degree this can be thought of as a natural part of validity and reliability; in fact, some have referred to it as system reliability. Nevertheless, given the demands placed on Title I assessments to detect change, especially from year to year, it becomes necessary to consider comparability in designing and developing the assessment, and then in gathering confirmatory data during the implementation phase. Although difficult to implement and to document, States have an obligation to show that they have made a reasonable effort to attain comparability, especially where locally-selected assessments are part of the system.

e. Procedures for test administration, scoring, data analysis, and reporting

Most States take great pains to ensure that the assessments are properly administered, that directions are followed, and that test security requirements are clearly specified and followed. Nevertheless, it is important they document the ways in which they ensure that their system does not omit any of these basics.

f. Interpretation and use

Although this topic is closely related to that of validity, and is discussed in most of the other topics in this section, it is mentioned here because of its importance. Even if an assessment is carefully designed, constructed and implemented, it all can come to naught if users are not helped to draw the most appropriate interpretations and to use the results in the most valid ways.

Technical quality and stages of development. Technical quality relates to the three main stages or phases of the development and implementation of an assessment system:

- Design and development
- Initial implementation
- On-going revision and improvement

These stages present opportunities both to ensure quality and to document that quality. Some of the elements are more related to the initial stages, some to the implementation/maintenance phases. A few implications are briefly mentioned below.

- At the **design/development stage,** the State has the best opportunity to focus on validity, reliability and fairness. This is the appropriate time to ensure that the assessment is aligned, that it is long enough to yield reliable scores, and that the items

[8] This is not to be confused with requirements of program or fiscal comparability.

give all students a fair opportunity to demonstrate their skills. These are standard assessment development processes, and States ought to be able to present substantial evidence of technical quality. Content validity is usually ensured by the development process, including the method of translating the standards into assessment blueprints or specifications, involving teachers and content specialists in the process, and on-going, systematic matching of the assessment items and tasks with the standards. This is also a way to document alignment. All of the issues discussed under the alignment section are relevant here, not the least of which is assuring that the standards which call for higher-order skills and understanding are adequately assessed.

- As part of the **initial implementation phase**, many States conduct studies to verify that the design principles actually produced an assessment with the qualities desired. For example, States can exploit the larger samples of student data to confirm the technical characteristics of the assessment tasks, including the fairness of the tasks for different student populations, and to confirm the link between the assessments and the performance standards. States should also be able to describe steps to ensure proper administration, scoring, analytic procedures, and reporting practices.

- During the **on-going annual administration** of the assessment program, if not before, States will want to confirm the proper use and interpretation of the results. This often leads to, and is done as part of, a statewide staff development effort focused on the use of the results to help teachers better identify and strengthen weaknesses in the instructional program.

Construct validation efforts continue throughout the life of the assessment. Evidence should continually be sought that the results truly reflect the goals of instruction, especially those related to higher-order thinking and understanding. In fact, with the spotlight of accountability on assessment results, it is all the more important to be sure that the assessment--which might not have changed at all--is still assessing the same skills. Under the pressures of accountability, steps taken to improve scores can change the natural relationships between instruction and assessment. Assessment items that ordinarily tap higher-order thinking skills might actually reflect more rote skills if certain types of test-preparation efforts are used. The unfortunate side-effects of accountability also make it advisable for States to document more basic aspects of quality, including the fact that students have not been deliberately taught the actual assessment tasks or clones of the items that would spuriously improve results.

Finally, it is obviously not possible to study the consequences of an assessment until it has been implemented for a year or more. One approach to this validation effort might be to pose a number of questions, then search for links to the assessment results. For example: Are more students meeting the standards because the results led to the creation of a dynamic statewide after-school program? Are more students being retained in grade as a result of the assessment results? Are more teachers part of a long-term professional development program that improves the teaching of reading to low-achieving students?

5. Preparing a State submission of evidence

Evidence to support the existence of quality for each of the six characteristics of technical quality may take many forms, including requests for proposals; technical manuals; instructions and materials associated with the assessments and the reports; professional development descriptions and materials; and other descriptive materials. Appendix D outlines some illustrative types and sources of evidence that peer reviewers might look for under each category of technical quality.

States are expected to present a persuasive body of evidence to support the quality of their assessments, including evidence about the quality of the assessment instruments themselves and evidence about how they are used. Although the States are expected to have some validation evidence in hand, in reality, validation requires the accumulation of information from many sources over time. Most States will be judged on the basis of the quality and thoughtfulness of their long-range plans for obtaining evidence showing that (1) the assessment instruments do in fact assess the intent of their standards, that (2) assessment information is interpreted and used properly, and that (3) unintended negative consequences are minimized. This scenario is consistent with the nature of technical quality--it is not a simple "have-not have" issue, but a process of continuous improvement and successive documentation over the years.

6. Questions for peer reviewers:

Peer reviewer questions	Desirable Evidence	Incomplete or Unacceptable Evidence
D1. How has the State considered the issue of **validity** (in addition to the alignment of the assessment with the content standards) and taken steps to ascertain that the assessments are measuring the knowledge and skills described in the standards--and that the interpretations are appropriate? Has the State specified the purposes for the assessments, delineating the types of uses and decisions most appropriate to each?	Peer reviewers will look for evidence of construct validity, consequential validity, and evidence that State and local users draw valid inferences from the assessments. They will want to see that the State took care in developing the assessment (meaning it conducted field tests and various types of research efforts) to be sure that the items and tasks actually tapped the essence of the standards--and that it did so for students of diverse backgrounds. In addition, they will want to see that the State has a systematic plan for conducting on-going validation studies to see if the results should be trusted. For example, it may want to compare the assessment results with other assessment information and/or with the quality of work that students	The State conducted an alignment study, but has no plans for studying the assessment to see if it actually assesses what it claims to, or to seek to identify types of students or schools where the results are not valid because for one reason or another the assessment does not function as it was designed.

Peer reviewer questions	Desirable Evidence	Incomplete or Unacceptable Evidence
	are actually producing in class. Moreover, validation studies should document the impact of the assessments. For example, they may want to see if the assessments have had a positive impact on classroom practice – e.g., whether they are doing more writing and thought-provoking project work, or whether they are spending an inordinate amount of time in lower-level test prep activities.	
D2. How comprehensively has the State determined that its assessments provide consistent and **reliable** results for individual students, schools, and LEAs? Does the State include information in its reports about the level of reliability of its scores?	Peer reviewers will look for evidence from the design of the test, analyses of test and scoring data, procedures for ensuring rater reliability, steps taken to ensure reliability of school-level scores, and communications and training opportunities for schools and the public to understand the level of reliability of the assessment. For example, one State required in its request for proposals for developing the assessment that the bidders provide evidence on the relative advantages and disadvantages of various test lengths and configurations, given the purposes for the different components of its assessment system. This information was examined by State staff and its technical advisory committee before making final decisions.	The State uses a short version of a standardized test not only for school-level assessment purposes but also for making important decisions about student promotion and placement.
D3. What steps has the State taken steps to ensure the **fairness and accessibility** of the assessments?	Peer reviewers will look for evidence that the State has taken steps to ensure fairness in the development of items and tasks, including the conduct of bias studies; in the administration of the test; and in the reporting of results. They will expect to see how accommodations and alternate assessments are used to help students respond to tasks in a meaningful fashion, as well as statewide figures on the numbers of students who used different accommodations and the achievement results for each group. States are expected to demonstrate that they assess a high percentage of	A large number of students are not assessed and the State has no clear plans for increasing the proportion assessed, or has no program for confirming that the results are valid for those students who are assessed.

Peer reviewer questions	Desirable Evidence	Incomplete or Unacceptable Evidence
	all students, that they have a solid plan for increasing that percentage, and that LEAs have an incentive for assessing as many students as possible.	
D4. How are **multiple measures** used to meet the criteria of validity, reliability, and fairness?	The State developed a matrix of the ways in which multiple measures might enhance the technical qualities of the assessments; this became a template that guided the initial design of the program and the assessment blueprint. The State can show how different measures and the use of different formats and strategies are used to increase the validity, reliability and fairness for each of its assessments for each of its population groups.	The State uses a single norm-referenced test and counts some items in more than one domain to accomplish coverage of all of the standards.
D5. In what way does the State ensure that the assessment results are **comparable** for different schools and for different years?	Peer reviewers will look for evidence of year-to-year consistency in development, administration, scoring, and analysis procedures, as well as evidence that the item content and focus and level of challenge are maintained from year to year, including the use of statistical procedures to link scores on different forms of the tests.	
D6. What evidence does the State have that its **administration, scoring, analysis, and reporting** procedures consistently meet high technical standards?	The State developed a set of criteria or standards for each of these components. It requires its contractors to provide specific information on the degree to which each criterion is met. This information is then reviewed by the State staff and appropriate advisory committees.	There are no procedures for ensuring that teachers or students do not have inappropriate access to the assessments. There are no procedures for ensuring that the scoring of open-ended tasks meets industry standards for accuracy.
D7. What actions has the State taken to ensure that teachers, other educators, and parents properly **interpret and use** the results? How does the State help them take into account the accuracy of the results when making interpretations?	The reports themselves contain considerable information and use graphics to aid proper understanding. The State routinely prepares and distributes brochures and manuals specifically designed for different audiences to help them interpret the results. These documents contain a variety of scenarios that illustrate different problems and issues. It also	The results are distributed with a minimum of supplementary information, including, for example, only a very brief definition of each of the figures in the report.

Peer reviewer questions	Desirable Evidence	Incomplete or Unacceptable Evidence
	conducts annual workshops for school personnel focusing on the different reports and how they can be used.	
D8. What steps is the State taking to periodically **review and improve** its assessments?	The State has several advisory committees that monitor different aspects of the assessment system and review the results of periodic studies of problem areas. The State's assessment statutes call for an objective evaluation every five years, and the State's assessment budget specifically provides for the conduct of evaluations and research studies, and for the ongoing upgrading of the assessments.	The State has no plan or procedure for improving the assessment.

PART III. Reporting and Using Assessment Results in Accountability

The last two sets of requirements pertain to the reporting of results of the assessment and the use of the results for determining adequate yearly progress of the schools. They are:

Reporting requirements:

E. Providing individual student reports
F. Providing disaggregated group reports
G. Development and dissemination of school performance profiles

Using assessment information for accountability purposes:

H. Ensuring that State assessment is the primary basis for determining adequate yearly progress (AYP)
I. Including students who have attended the same school for a full academic year

Part III - E: Providing Individual Reports

1. Requirement-Legal Citation:

> State assessments shall –
> Provide individual student interpretive and descriptive reports, which shall include scores, or other information on the attainment of student performance standards.
> (Sec. 1111(b)(3)(H))

2. Intent and Purpose

The intent of this requirement is to ensure that some level of individual student reporting is available as part of the assessment so that students, teachers, and parents have access to information about individual student performance. Learning is essentially an individual matter; improving performance without feedback is inefficient at best, and hopeless at worst.

3. Description

The statutory requirement does not specify how extensive or detailed individual student reports need to be. However, it is important that individual student data be reported in relation to the State's content and performance standards.

Some State assessments are using matrix sampling procedures that are not designed to provide complete data on each student since no student takes the entire exam. Such States must provide some level of student reporting either on the portions of the test taken, or from other sources of information that relate to the State's standards.

4. Preparing a State submission of evidence

States should provide examples of student reports, descriptions of the types of information that the reports include, the sources of the data on the reports, the general ways in which the results are presented, the frequency and timeliness of the reports, the ways in which various types of reports are used to inform parents of their children's progress, how the reports are used by school personnel to improve programs, and how all users are trained to properly interpret the findings.

5. Peer Reviewer Questions

Peer reviewer questions	Desirable Evidence	Incomplete or Unacceptable Evidence
E1. How does the State provide individual student reports? What is the source of the data?	The State provides individual information from the State assessment, or it requires that LEAs report the results of other assessments.	The State provides student reports with course grades that do not relate that information to the State's content and performance standards.
E2. What is contained in the student reports? How are the data presented? Are the results based on the State's content and performance standards?	The reports indicate how well each student has performed relative to the content and performance standards, using both narrative and graphic modes.	Student reports are based on a matrix-sampling design that provides information on some parts of some standards, with no provision for reporting on the other standards. Reports only give composite national percentile ranks without linking the results to the standards.
E3. How does the State ensure the quality of these reports?	The State monitors the quality of all contractor-produced reports using State assessment information, and/or annually monitors the quality of LEA- produced reports against criteria that have been developed and disseminated.	
E4. How are the results disseminated and communicated? Are they clear and understandable?	A description of strategies to ensure that individual reports go to all parents in understandable ways; that ensure that parents can see how their children do in relation to the standards; and that the reports show how much students have progressed since the last assessment.	
E5. How is the State supporting the appropriate interpretation and use of the student level reports?	The State produces interpretive guidelines and manuals. The State conducts training for local personnel in ways to improve usefulness of individual reports. The reports describe the amount of error that is associated with each score.	Reports imply that the results are without error.

Part III – F: Disaggregated Reporting

1. Requirement-Legal Citation

> The State assessments shall –
> Enable results to be disaggregated within each State, local educational agency, and school by gender, by each major racial and ethnic group, by English proficiency status, by migrant status, by students with disabilities as compared to nondisabled students, and by economically disadvantaged students as compared to students who are not economically disadvantaged. (Sec. 1111(b) (3)(I))

1. Intent and Purpose

The purpose of this requirement is to ensure that the progress of all student populations is annually and systematically monitored. This is a critical step in ensuring that all students are meeting challenging standards.

2. Description

States are required to provide State assessment data that are disaggregated for a variety of student subgroups in all schools and LEAs. States are required to provide for the reporting of results for a variety of student subgroups in all schools and LEAs, if the data are statistically sound. These data must be included in a public report on school progress that can be produced either by the State or by the school district.

Peer reviewers should look for descriptions of the nature of the reports, the procedures for distributing them, procedures for protecting student confidentiality, procedures for guarding against over-interpretation of small differences--especially for small schools and small subgroups, and for implementing professional development strategies for helping teachers and administrators interpret and use the results to improve programs. The Council of Chief State School Officers has produced a document that is designed to assist state and local personnel in carrying out this requirement.[9]

States should submit samples of their public reports – produced at either the State or local level. The narrative provided by the State should answer the questions provided below and explain the nature of the reports, how they are produced and disseminated, and how they are used.[10]

3. Preparing a State submission of evidence

---

[9] One of the inevitable questions is, "How large should a group of students be before we report the results?" Jaeger and Tucker (1998) advocate the increasingly common practice of not less than ten students in a single group.

[10] The Council of Chief State School Officers has produced a document that is designed to assist state and local personnel in doing exactly this. Jaeger, R.M & Tucker, C.G. (1998) Analyzing, disaggregating, reporting, and interpreting students' achievement test results: A guide to practice for Title I and beyond. Washington, DC: Council of Chief State School Officers.

States should provide reports that demonstrate how their data are disaggregated. A description of the categories and rules for disaggregating data with small numbers would also be useful.

4. Peer Reviewer Questions

F1. Which disaggregated student achievement results are reported at which levels? (By grade level and content area, as appropriate)	Gender	Racial & ethnic groups	English proficiency status	Migrant	Disabled vs. non-disabled	Economically disadvantaged vs. non-disadvantaged
School						
LEA						
State						

Peer reviewer questions	Desirable Evidence	Incomplete or Unacceptable Evidence
F2. If all levels of the reports are not produced by the State, how does the State confirm that locally developed reports are produced and disseminated?	States routinely collect copies of all locally-produced reports, or monitor them as part of a systematic program quality review process.	The State produces disaggregated reports for state-level information, and informs LEAs that they are obligated to do the same, but no monitoring is done.
F3. How are public reports disseminated?	Peer reviewers should look for systematic procedures for annually reporting these results, either as part of the general reporting, or as part of a special process.	There are no policies or procedures to ensure that they are disseminated.
F4. What are the State policies regarding reporting results for small schools and small student subgroups? How does the State ensure that LEA and school personnel do not over-interpret the findings? Is student confidentiality ensured?	Peer reviewers should look for evidence from policies on reporting. In general, groups smaller than 10 students are not recommended for such reporting. Peer reviewers might also consider evidence of multiple measures that are used to increase the validity and usefulness of reporting for various student groups. The State disseminates special interpretive reports for small schools, and/or conducts training on ways to ensure sound interpretation and use of results.	No guidance is provided, or it is too general to be useful.
F5. How does the State use disaggregated information to ensure that statewide policies and	Peer reviewers should also look for comments in various documents that the State has gone	

Peer reviewer questions	Desirable Evidence	Incomplete or Unacceptable Evidence
procedures regarding curriculum and other aspects of their reform program are reinforcing the importance of all students mastering the standards? How does the State help LEAs do the same?	the "extra mile", i.e., that it is not only reporting disaggregated information but is taking steps to make sure that the data are used. For example, the State might show that it is requiring CSRD applicants to show that they have systematic plans to use disaggregated information.	

Part III – G: Development of District and School Profiles

1. Requirement-Legal Citation

Each LEA shall publicize and disseminate to teachers and other staff, parents, students, and the community, the results of the annual review of all schools served under Title I in individual school performance profiles that include statistically sound disaggregated results. (section 1116(a)(3))

2. Intent and Purpose

The intent of this requirement is to make sure that the public as well as all school personnel are aware of the progress of students in the school in meeting the State's content and performance standards. In order to understand that progress and to help set the direction for continued improvement, it is important that the whole school community also see the school's areas of relative strength and weakness, as well as other information about the school's population, its programs and its resources.

3. Description

Title I requires that all participating LEAs produce individual school performance profiles for all of their participating schools. The law requires that the profiles--

- include the results of the LEA's review of the progress of all schools served under Title I, based on State assessments and any additional local measures that are used to determine whether a school is making adequate yearly progress;
- include statistically sound data disaggregated by the groups listed in "F" above; and
- be publicized and disseminated to teachers and other staff, and to parents, students and the community.

The law also permits LEAs to include other appropriate information in the profiles, such as data on teachers' qualifications; class size; and attendance, promotion rates and retention rates.

4. Preparing a State submission of evidence

Sample school profiles should be accompanied by a narrative description of how district and school profiles are produced and disseminated, how quality is ensured, and what is included.

5. Peer Reviewer Questions

Peer reviewer questions	Desirable Evidence	Incomplete or Unacceptable Evidence
G1. Do all participating LEAs annually develop and disseminate performance profiles for all their schools that receive Title I funds? How does the State ensure that they do and that they contain all the required information?	The State has collected copies from all LEAs and confirmed that they contain both the AYP information and the results disaggregated by the required student groups. The State provides a work plan and timeline for development and dissemination of profiles for every district and school in 2001.	The State has a policy that requires that LEAs develop and disseminate the profiles, and it has informed the LEAs of that policy.
G2. What does the State do to assist LEAs in producing profiles that are of high quality and are useful in improving school programs?	The State has developed and disseminated guidelines and model profiles and held workshops across the State. The State has an active web page devoted to the development and improvement of school profiles.	
G3. How does the State document that LEAs publicize and disseminate the profiles to all the required audiences?	The State requires that one of the means of dissemination be the internet and it surveys the reports that are posted. The State conducts annual surveys of its citizens for other purposes, and it asks how many have seen the school profiles.	

Part III - H. Ensuring that State Assessments are the Primary Basis for Determining LEA and School Progress

1. Requirement--Legal Citation

> Each State plan shall demonstrate that the State has developed or adopted a set of high quality, yearly student assessments ... that will be used as the primary means of determining the adequate yearly performance of each local educational agency and school served by this part. (Sec. 1111(b)(3))

2. Intent and Purpose

The requirement that States establish State assessment systems is to ensure that all students are held to challenging standards and that schools are held accountable for ensuring that every student meets the State's standards. In the 1994 reauthorization of the Elementary and Secondary Education Act, States are to establish measures for adequate yearly progress for local educational agencies and schools. These measures for AYP must be primarily based on the State assessment, though the use of other indicators of school performance, including non-academic factors such as attendance rates or graduation rates, is also permitted.

3. Discussion

The Title I statute provides for the use of non-academic factors in a State's definition of adequate yearly progress, but it is clear that State assessment information must be the dominant factor. The State must explain how assessment results are used for holding schools and districts accountable. In particular, the description of the States' accountability measures should demonstrate that assessment results account for most of the weight in any type of total index or composite. If the system does not produce some type of composite, the description needs to describe how State assessment functions as the primary factor among the State's measures.

4. Preparing a State submission of evidence

States should provide their definitions of adequate yearly progress for districts and for schools. Those definitions should be accompanied by an explanation of the factors that are considered in determining adequate yearly progress and the weights that those factors receive. Issues such as the timeframe for all students reaching State standards should also be discussed and explained.

5. Peer Reviewer Questions:

Peer reviewer questions	Desirable Evidence	Incomplete or Unacceptable Evidence
H1. In what way is student performance on State assessments defined as the primary element in the State's definition of adequate yearly progress for schools and districts?	The State's adequate yearly progress (AYP) index puts 80% of the weight on the average assessment score (averaged across grades and content areas), 15% on dropout rate and 5% on average attendance. (The percentages are purely illustrative!!) The State uses the State assessment results as the first screen for identifying program improvement schools, then permits LEAs to use local data, either local assessment results or non-cognitive factors, to refine the identification process. The State provides specific criteria for this process and monitors the use of the criteria to ensure that the intent of the law is not violated.	The State's (AYP) index gives 60% of the weight to the State assessment results and 40% to the local drop-out rate. The State does not control for the variances of the different measures, a practice that results in a higher weight being placed on dropout changes than assessment score changes.
H2. What role do local assessments play in defining AYP? Are they part of the "State's assessment system" or are they considered supplemental? If they are part of the definition for AYP, what steps are taken to ensure that they are of high quality?	The State's system encourages the development of local assessment systems. The results do not count in defining AYP, but they are used in the ongoing monitoring of student progress throughout the year, or in assessing grades and content areas not included in the state assessment system. The State's assessment system has defined local assessments in speaking and listening as part of the system. LEA personnel are thoroughly trained in the use of state-provided assessments and in specific procedures for common administration and scoring. The results are collected and closely monitored for technical quality. The results for a school can count as much as 20% depending on the	The State permits LEAs to use their locally-developed assessments to count as much as 10% in determining the AYP rate for a school with virtually no monitoring for technical quality.

Peer reviewer questions	Desirable Evidence	Incomplete or Unacceptable Evidence
	number of grade levels assessed. (The percentage is illustrative only!!)	
H3. If non-cognitive measures are used as part of the AYP definition, how are they weighted? Are they included in an index, or are they used as a secondary screen or filter?	The State's AYP system includes both attendance and dropout rate, but only as a "tie-breaker" for schools that are on the borderline of meeting the State's "program improvement" criteria—and then according to specific criteria and procedures.	The State computes an achievement index that takes poverty into account, such that a school with 80% poverty rate receives a higher index than a school with 40% poverty The State permits LEAs to determine the weight for a given non-cognitive factor The State encourages LEAs to add additional non-cognitive factors in applying the AYP criterion

Part III - I. Include students who have attended school in the LEA for a full academic year

1. Requirement--Legal Citation

> State assessment systems shall –
> Include, for determining the progress of the LEA only, students who have attended schools in the LEA for a full academic year, but who have not attended a single school in the LEA for a full academic year. (34 CFR 200.4(b)(8) – the regulations clarifying Sec. 1111(b)(3)(G))

The regulations state this requirement a bit more clearly:

> "(8) Include, for determining the progress of the LEA only, students who have attended schools in the LEA for a full academic year, but who have not attended a single school in the LEA for a full academic year."

2. Intent and Purpose

The intent of this requirement is to ensure that LEAs and schools are held accountable for those students that they have been educating for the current academic year. The provision recognizes that students may move among schools within a school district during the year. It therefore does not hold a school accountable for the performance of students who have not attended the school for a full academic year, but the district is still responsible for ensuring that such students are taught to challenging State standards. Therefore all students who have been in the LEA for a school year must be included in the LEA's accountability rating.

3. Preparing a State submission of evidence

States should describe which students and schools are counted in their determinations of adequate yearly progress.

4. Peer Reviewer Questions

Peer reviewer questions	Desirable Evidence	Incomplete or Unacceptable Evidence
I1. Has the State clearly informed the LEAs regarding which students must be considered in determining adequate yearly progress?	Peer reviewers will look for evidence that the State counts all students who have been in an LEA for a full academic year in that LEA's accountability rating. Mobile students do not need to be included in a school's ratings if they have not been in the same school for a full academic year.	The State does not collect information on student mobility as part of its State assessment. It does collect very similar information as part of its enrollment and attendance information collection procedures, but these data cannot be integrated

		with the student performance information.
12. Does the State make any effort to ensure that LEAs are following this policy?	The State monitors the degree to which LEAs are observing this requirement as part of its general compliance and review processes.	The State system allows for the use of local assessments, but the State is not able to ensure that LEAs include all appropriate students in their calculations.

References

American Educational Research Association, American Psychological Association, National Council on Measurement in Education. (1985). *Standards for educational and psychological testing*. Washington, DC: Author.

American Educational Research Association, American Psychological Association, National Council on Measurement in Education. (In press). *Standards for educational and psychological testing*. Washington, DC: Author.

Blank, R. & Webb, N. (February 1999). *Analysis of alignment of state standards and assessments in mathematics and science*. Council of Chief State School Officers and the National Institute for Science Education.

Hansche, L. N. (1998). *Handbook for the development of performance standards*. Washington, DC: U. S. Department of Education and the Council of Chief State School Officers.

LaMarca, P., Redfield, D., and Winter, P. (1999). *State Standards and State Assessment Systems: A guide to alignment*. A paper prepared for the Council of Chief School Officers SCASS group on Comprehensive Assessment Systems-Title I.

Messick, S. (1989). Validity. In R.L. Linn, R.L. (Ed.). *Educational measurement*, (3rd Edition, pp. 13-103). New York: Macmillan.

Messick, S. (1988). The once and future issues of validity: assessing the meaning and consequences of measurement. In Howard Wainer and Henry Braun (Eds.), *Test validity* (pp. 33-48). Hillsdale, NJ: Erlbaum.

National Center on Educational Outcomes. (1997). *Policy Directions 7*.

Rogosa, D. R. (1999) *Accuracy of individual scores expressed in percentile ranks: Classical Test theory calculations*. CSE Technical Report 509. CRESST.

Rogosa, D. R. (1999) *Accuracy of Year-1, Year-2 comparisons using individual percentile rank scores: Classical test theory calculations*. CSE Technical Report 510. CRESST.

Rogosa, D.R. (April 1994). *Misclassification in student performance categories*. Appendix to CLAS Technical Report. Monterey, CA: CTB/McGraw-Hill.

Webb, N. L. (1997). *Research Monograph No. 8: Criteria for alignment of expectations and assessments in mathematics and science education*. Washington, DC: Council of Chief State School Officers.

Webb, N.L. (1999). *Summary report: alignment analysis of standards and assessments for four states in science and mathematics*. Washington, DC: National Institute for Science Education & Council of Chief State School Officers.

Young, M. J., and Yoon, B. (1998). *Estimating the consistency and accuracy of classifications in a standards-referenced assessment.* CSE Technical Report 475. Center for the Study of Evaluation, National Center for Research on Evaluation, Standards, and Student Testing. University of California, Los Angeles.

Appendix A

Including LEP Students in State Assessments under Title I To the "Extent Practicable"

When the Elementary and Secondary Education Act was reauthorized in 1994, it launched a major sea change in how Federal programs support State and local education reform efforts. Central to the reforms embodied in ESEA is support for States to establish challenging content and student performance standards that apply to all students. No longer would some groups of students be condemned to low expectations merely because they happen to attend a high poverty school or because they are just mastering the English language. Rather, the reforms envisioned in ESEA reflect a belief that all students can learn to high standards and that schools and districts should be held accountable for ensuring that all of their students achieve at high levels. One of the most challenging issues in carrying out these reform efforts is appropriately including English language learners (referred to in the law, and therefore throughout this document, as limited English proficient (LEP) students) in a standards-based system and assessing their progress in meaningful, valid, and reliable ways.

Most of the requirements related to establishing standards, assessments, and accountability systems are contained in the statutory language of Title I of ESEA. Title I requires each State to establish challenging content and student performance standards and a set of high-quality, yearly assessments aligned with those standards by the 2000-01 school year. These standards and assessments are meant to ensure that all students are held to high expectations. Although the statutory language applies specifically to Title I schools, it clearly specifies that if a State is developing statewide standards and assessments, then those same standards and assessments should apply to Title I schools. To date, no State is pursuing a standards and assessment approach that applies only to Title I schools.

The Title I language requires States to develop statewide assessments that include a variety of elements necessary to have a meaningful system. Such assessments must include at least reading and mathematics in at least three gradespans. They must be technically sound, aligned to standards, include multiple measures, and be used as the primary means for holding schools and districts accountable. These assessments must also be used to measure the performance of all students in valid and reliable ways so that schools can be held accountable for their performance. The Title I assessment requirements have specific language related to promoting the full inclusion of all students:

> State assessments shall provide for –
> (i) the participation in such assessments of all students in the grades being assessed;
> (ii) the reasonable adaptations and accommodations for students with diverse learning needs, necessary to measure the achievement of such students relative to State content standards; and
> (iii) the inclusion of limited English proficient students who shall be assessed, to the extent practicable, in the language and form most likely to yield accurate and reliable

information on what such students know and can do, to determine such students' mastery of skills in subjects other than English. (Section 1111(b)(3)(F); 34 C.F.R. 200.4(b)(7))

The intent of these requirements is to: 1) ensure that all students are held to the same high standards and appropriately assessed against those standards; and 2) ensure that all students are part of the indicators used to hold schools accountable. Including LEP students in assessments in valid and reliable ways is already required under Title VI of the Civil Rights Act, and Title I provides some flexibility for States to use a variety of strategies to make certain that all students participate in the assessment system. For example, such strategies as accommodations, alternate assessments, assessments in the students' primary languages, and plain language assessments may be part of the State's system.

It is clear that interpreting and implementing the statutory language that refers to assessing LEP students "to the extent practicable, in the language and form most likely to yield accurate and reliable information on what such students know and can do, to determine such students' mastery of skills in subjects other than English" is a complex task. The question of how to include LEP students in the State's assessment has no single answer. Historically, most conventional tests have been normed on native English speakers. As a result, assessments that have not been designed to include LEP students may not yield accurate and reliable information about what LEP students know and can do. These assessments may in effect be measuring English language skills rather than the knowledge and skills in other content areas for which the assessment was intended.

The purpose of this guidance is to begin to help clarify what is meant by the Title I assessment requirements to include LEP students. It has been developed in consultation with assessment directors, education organizations, and State and local policymakers, a process that has taken a long time given the complex nature of this subject and the fact that the field continues to develop knowledge in these areas. The guidance has three parts:

I. **State and district considerations**: The first part of this paper presents a series of questions that States, districts, and schools should consider when determining the most appropriate methods for including LEP students in the assessment and accountability systems. After considering the following questions, States and districts should be able to determine the most appropriate strategies for including LEP students and direct resources and expertise to meet these challenges and ensure that all students reach challenging State standards.

II. **Scenarios**: The second part of the guidance includes a few hypothetical scenarios that illustrate implementation of the requirements.

III. **Resources**: The last part of the guidance presents additional resources that States, districts and schools may wish to utilize for maximizing inclusion of all students.

I. STATE AND DISTRICT CONSIDERATIONS

Determining the most appropriate and most feasible assessment strategies for LEP students must be considered within the context of each State or district. A one-size-fits-all approach is impossible given the variety of student populations, their language proficiency and concentration, and the range of instructional programs that are offered across the States. Therefore each State should consider the following questions when determining whether it is assessing LEP students, to the extent practicable, in the language and form most likely to yield accurate and reliable results.

Who are the LEP students in the State or district? What languages do they speak?

LEP students are national origin minority students who cannot speak, read, write or comprehend English well enough to participate meaningfully in and benefit from the school's regular education program. School districts have an obligation to provide instructional services that will enable these students to overcome language barriers to academic achievement.

The first step in determining strategies to include LEP students in the State assessment system is to analyze the LEP student population in the State in order to decide what language and form of assessments will yield accurate and reliable information about what they know and can do. States and districts should consider what languages their LEP students speak, their native language literacy, their language of instruction, and their level of English language proficiency. States should also consider their demographic trends to begin to predict the needs of future populations as best as possible.

What is the instructional approach or language of instruction for LEP students in the State or district?

The instructional approach and language of instruction used to teach LEP students should be considered in determining how to assess those students. Native language assessment may be appropriate if a student is receiving instruction in his/her native language or if a student can better demonstrate his/her course content knowledge through his/her native language, regardless of the language of instruction. Native language assessment might not be appropriate, however, if a student has never received instruction in his/her native language and lacks literacy skills in that language. If a child is in a specially designed English instructional program for LEP students, an English language assessment may be more appropriate, particularly with accommodations that reflect the special instructional strategies and approaches used in the classroom.

Would assessments (in subjects other than English) yield more accurate and reliable information if the tests were in the English language or in a native language?

The language and form of assessments most likely to yield accurate and reliable information on LEP students' performance are dependent on such factors as English proficiency level, native language literacy and proficiency, and type of instructional program.

Using English language versions of tests: States and districts should consider whether an assessment in the English language is the one most likely to yield accurate and reliable information on what a LEP student knows and can do, or whether a native language assessment is more appropriate. Among the factors to consider in making such determinations are the student's English language proficiency in the four domains (speaking, listening, reading, and writing) and the number of years the student has received academic instruction in English.

Appropriate adaptations and accommodations may be needed to facilitate inclusion of LEP students in the State or district assessments that are administered in English. Examples of adaptations or accommodations that States and districts have used include extended time, flexible scheduling, small group administration, simplified directions, larger print, audiotaped instructions or questions, use of bilingual dictionaries and native language glossaries, audiotape responses, and separate testing sessions for LEP students. Because of the diversity within the LEP student population, no single method would be likely to be effective for all LEP students. For this reason, providing a range of adaptations or accommodations is important to achieving the goal of providing accurate and reliable information about what students know and can do. It is also important to ensure that validity and reliability of the assessment instrument are not compromised if adaptations or accommodations are used.

Using native language versions of tests: Even with accommodations, there may be LEP students for whom a test in English is not the one most likely to yield accurate information on what they know and can do. These students must be assessed, to the extent practicable, in the students' native languages in order to produce accurate and reliable information on what the students know and can do to determine their mastery of skills in subjects other than English. It should be re-stated, however, that native language assessment might not be appropriate if a student has never received instruction in his/her native language and lacks literacy skills in that language. If a child is in a specially designed English instructional program for LEP students, an English language assessment may be more appropriate, particularly with accommodations that reflect the special instructional strategies and approaches used in the classroom.

Whether it is practicable to assess students with a separate valid assessment in each native language depends on a number of considerations (e.g., the language of instruction, the alignment of the test to the State's standards, the number of students who speak a given language, the students' proficiency in their native language, and the appropriateness of commercially available native language assessments). For example, if a State has a large number of LEP children whose native language is the same--e.g., Spanish--it would likely be practical, and thus required, for the State to assess those students in their native language if that is the most appropriate measure of the knowledge and skills of those students. Indeed, in most States, the population of Spanish-speaking students is large enough to justify the development of Spanish versions of the assessments. Such assessments would, of course, need to be aligned with the State's content and performance standards. Each State or district must examine its student population and determine how best to include LEP students.

If separate native language assessments are not practical, a State and district are required to use

other measures to assess LEP students' progress. Those measures may include classroom performance measures such as portfolios, student progress reports, teacher observation checklists, student performance evaluations, teacher-student conference interviews, and an annual English language proficiency review. Regardless of what instruments are used to assess LEP students, those instruments must cover the same standards being assessed for all students.

Has the State or district met civil rights legal requirements relating to assessing LEP students?

School districts have a responsibility to comply with Title VI of the Civil Rights Act of 1964 (Title VI) (42 U.S.C. 2000d; 34 C.F.R. 100.3), which prohibits States and school districts that receive Federal financial assistance from excluding from participation, denying benefits to, or discriminating against students, on the basis of race, color, or national origin, in all of their operations. Under Title VI, which is enforced by the Department's Office for Civil Rights (OCR), States and school districts have a responsibility to provide equal educational opportunity to students who are limited English proficient. In Lau v. Nichols, 414 U.S. 563 (1974), the Supreme Court found that failure to provide equal educational opportunity to language minority students violates Title VI.[1] Under Title VI, school districts have an obligation to provide LEP students with alternative language services to enable them to acquire proficiency in English and to provide them with meaningful access to the content of the educational curriculum that is available to all students.

To the extent statewide or district assessments may result in an educational benefit to non-LEP students, the exclusion of LEP students from participation in statewide assessments or a failure to provide LEP students with accommodations may raise Title VI issues. Also, under Title VI, if LEP students are excluded from a particular statewide or district assessment based on an educational or psychometric justification, then districts have an obligation to collect comparable information about these students' academic progress.[2]

States and school districts are also subject to the Equal Protection Clause of the Fourteenth Amendment, which prohibits discrimination on the basis of national origin. The Equal Educational Opportunities Act (EEOA), (20 U.S.C. 1703(f)) prohibits States and school districts, among other things, from denying equal educational opportunities to an individual because of his or her national origin due to the failure to take appropriate action to overcome language barriers

---

[1] In Lau, the Supreme Court upheld a Health, Education, and Welfare (the predecessor to the Department of Education) May 25, 1970 policy memorandum, which advised school districts of their responsibility under Title VI to provide equal educational opportunity to national-origin minority students who are limited English proficient. The Court determined that, under Title VI, where the inability to speak and understand the English language excludes such students from effective participation in a district's educational programs, a district must take affirmative steps to rectify the language deficiency in order to open its instructional program to these students.

[2] If States or school districts use assessments under Title I for high-stakes decisions that affect individual students, such as student promotion or graduation decisions, Title VI requirements would also apply to that use of the tests.

that impede equal participation by students. Individual States or school districts may also be subject to other legal requirements that relate to the participation of LEP students in standardized assessments.

Has the State or district considered a variety of assessment options?

A State has considerable flexibility in designing its assessment system, provided the assessments are aligned with the State's content and student performance standards. For example, the State might select commercially available tests, develop its own tests, or opt for a combination. If a State selects a commercially available standardized test, inclusion of LEP students should be taken into account. Is the test available in other languages? What are the effects on reliability and validity of test scores for LEP students if accommodations are used? On what populations was the test normed?

Criterion-referenced tests, depending on their specific characteristics and psychometric properties might lend themselves to accommodations during testing that do not compromise the validity of the scores. A State that is developing its own assessment system should include LEP students in the design, piloting, and field-testing.

A State might consider joining one or more other States in order to develop native language assessments. This strategy allows for leveraging various resources that might not be otherwise available, and also provides an opportunity to learn from the expertise of other organizations that are involved in similar efforts. It would require working through challenges such as alignment with standards from different States and ensuring LEP populations from all States are included, but it may prove a smart choice for States with particular populations of LEP students.

Has the State or district utilized all available resources to ensure that LEP students are assessed to the extent practicable in the language and form most likely to yield accurate and reliable results?

A wide variety of resources are available to help a State or district include LEP students in its assessment system. For example, funds under Title I, Title VII, Migrant, and Immigrant Education programs may be used to include LEP students in assessment systems. Federal technical assistance providers may also serve as resources. In addition, State and local funds, such as those from State/local bilingual education programs, may be available. Working with test publishers to incorporate specific requests or requirements in their tests is another way of addressing the inclusion of LEP students. States and districts may also request assistance from the U.S. Department of Education to interpret requirements and access technical assistance on these issues. Given the range of resources available, every State should be able to take proactive steps toward fully including LEP students in the State assessment system in appropriate and meaningful ways.

II. SCENARIOS

Would a State that has an English-only law be in compliance with the Title I assessment requirements?

Title I requires States to include LEP students in final assessment systems used to hold schools and districts accountable for student performance. This can only be done in a meaningful way if such students are tested in ways that produce valid inferences about the progress of LEP students toward State standards. Toward this end, Title I makes clear that for subjects other than English, if native-language assessment of LEP students is practicable, and if it is the form of assessment most likely to produce valid information on their academic performance, then States *must* utilize such assessments. Thus, a statewide English-only assessment policy would conform to Title I only if the State could demonstrate, for subjects other than English, that no native-language assessment of LEP students is practicable or that some other form of assessment is more likely to produce valid information on what LEP students know and can do.

If a State can show that native-language assessment is inappropriate or impracticable, Title I still requires the State to provide reasonable adaptations and accommodations necessary to measure the achievement of LEP students relative to State standards. Also, the State may supplement its statewide test with other measures that provide meaningful information about the performance of LEP students.

If the LEP students in my State speak numerous languages, none of which is most predominant, must native language assessments be provided?

This determination would need to be considered within the context of the types of instructional programs offered in the State and the level of English language and native language proficiency within the student population. If there are no concentrations of LEP students with the same language, then they most likely do not have instruction in their native language. So testing in English may be appropriate. However, accommodations or adaptations may be necessary to ensure that such students are provided an opportunity to demonstrate what they know and can do in a given subject. This can generally be determined given the types of instructional programs such students are receiving and the accommodations and adaptations that are provided to them in that context.

Instead of the statewide assessment given to students proficient in English, may a State test LEP students with a different test commercially available in another language, such as Spanish?

If another test is used for holding schools and districts accountable for the performance of LEP students, then the State must ensure that the test is aligned to the State standards and that its performance standards for LEP students represent the same level of knowledge and skills and rigor as the performance standards that are tied to the statewide test.

Must States have native language assessments available in subjects other than English for secondary school students who are new arrivals to the U.S.?

Schools receiving new secondary school students who do not speak English must first determine the native language skills of those students in order to determine how to most effectively address their educational needs. Clearly, a student who is literate in his or her native language will not require the same instructional program as a peer who is not literate.

Some of these students will only be able to accurately and reliably demonstrate what they know and can do in their native language. In such cases, native language assessments should be made available for subjects other than English.

III. Resources

There are a growing number of publications and research that deal with how to include LEP students in assessment systems. They cover issues of system design as well as very specific issues such as test translation. Technical assistance providers, such as the Department's Comprehensive Centers, can also provide information on what the literature says or what else has been done on this topic. Inclusion efforts in other States and districts can also help inform this issue. Finally, test publishers can contribute information. Many have been working extensively on assessment of LEP students, and their efforts can help inform this process within a State's specific context.

Some specific resources include:

Council for Chief State School Officers (CCSSO)
One Massachusetts Ave., NW Suite 700
Washington, DC 20001-1431
Contact: Julia Lara or Wayne Martin
http://www.CCSSO.org
(202) 408-5505
For information on standards, assessments and State Collaborative on Assessing Student Standards (SCASS) contact John Olsen.

Language and Diversity Laboratory Network Program (LCD LNP)

The Program is led by the three regional laboratories that share the language and diversity specialty areas: the Northeast and Islands Laboratory at Brown (LAB), the Pacific Resources for Education and Learning (PREL) and the Southwest Education Development Laboratory (SEDL): http://www.sedl.org/culture/lnp.html
The three laboratories pool their efforts for this project. The program intent is to help educators deal with issues of language and cultural diversity through professional development. This involves educating teachers about up-to-date work in relevant fields.

creating environments in which they can learn from one another, and helping them apply the best of current theory and practice to their particular situations.

The National Center on Educational Outcomes (NCEO) was established in 1990 to provide national leadership in the identification of outcomes, indicators, and assessments to monitor educational results for all students, including students with disabilities.
Http://www.coled.umn.edu/NCEO/
Director: Jim Ysseldyke, Ph.D.; Associate Director: Martha Thurlow, Ph.D.
National Center on Education Outcomes
University of Minnesota
350 Elliott Hall
75 East River Road
Minneapolis, MN 55455
Voice: (612) 626-1530
Fax: (612) 624-0879

National Clearinghouse for Bilingual Education (NCBE) is funded by the Office of Bilingual Education and Minority Languages Affairs (OBEMLA). NCBE provides practitioners with information on the education of limited-English-proficient students. NCBE compiles information on materials, programs, research, and other resources that can help educators meet the challenge posed by the complex and changing educational needs of language minority students in U.S. schools.

Center for Research on Education, Diversity & Excellence (CREDE) assists the nation's diverse students at risk of educational failure to achieve academic excellence. Central to its mission, CREDE's research and development focus is on critical issues in the education of linguistic and cultural minority students and those placed at risk by factors of race, poverty, and geographic location.
University of California, Santa Cruz
1156 High Street
Santa Cruz, CA 95064
(408) 459-3500
Director: Roland Tharp
http://www.crede.ucsc.edu
OERI Contact: Gilbert N. Garcia (202) 219-2144

Center for Evaluation, Research, Standards and Student Testing (CRESST) at the University of California, Los Angeles. CRESST conducts research on important topics related to K-12 educational testing. http://cresst96.cse.ucla.edu

Co-directors: Eva L. Baker and Robert Linn
Graduate School of Education
1339 Moore Hall
405 Hilgard Avenue

Los Angeles, CA 90024
(310) 206-1530

National Research and Development Center for English Learning and Achievement (CELA)

The Center on English Learning and Achievement (CELA) is dedicated to improving the teaching and learning of English and language arts. CELA's research seeks to learn what elements of curriculum, instruction, and assessment are essential to developing high literacy and how schools can best help students achieve success. We provide that information to teachers, schools, and communities so that they can choose the approaches that will work with their students. Our research is also designed to examine the tradeoffs (including costs) involved in using different approaches to English achievement. In short, our activities are planned to provide definitive information about what works, for whom, and under what conditions. CELA serves as the National Research Center on Student Learning & Achievement in English and is funded primarily by the U.S. Department of Education.
http://www.albany.edu/cela/
Co-directors: Judith Langer, Arthur Applebee, and Martin Nystrand
University of Albany, State University of New York
School of Education
1400 Washington Avenue
Albany, NY 12222
(518) 442-5026

Harvard Projects in Active Cultural Engagement (PACE)

In collaborating with schools in the area of cultural engagement, PACE: develops curriculum and performance assessment methods that engage young people in the process of defining excellence; designs technological tools for promoting active cultural engagement; creates sustained partnerships with arts and cultural organizations; and publishes research on the importance of the arts and humanities in promoting active cultural engagement.
http://www.pace.harvard.edu
Director: Dennie Palmer Wolf
18 Story Street
Cambridge, MA 02138
(617) 496-2770

Appendix B

Must all the standards be assessed?

Yes. All content standards that have been approved by the State as essential knowledge or skills for all students should be assessed. States may choose to develop uniform statewide assessments that address all standards, or they may adopt a model that reserves responsibility for assessing some standards to the State while assigning responsibility for assessing others to the LEA or school. For all assessments used for Title I accountability, the State must assure appropriate technical quality.

Figure 3 displays three levels of analysis for assessment results based on state and local assessments. Each level uses data from the same assessments to answer different questions or support different kinds of decisions.

- State-level analyses. The focus is strictly on questions that are relevant at the State level, such as, "What proportion of the students in grade four are able to read at the proficient level?" (Yes, it would include the question, "What percent of the schools are in need of improvement? but the process for making the decision about a school is seen as a school-level issue and is discussed under School/district-level analyses below.)

- School/district-level analyses. This perspective includes a wide range of uses for building- and LEA-level data, but focuses almost exclusively on the identification of schools that are in need of improvement under Title I.

- Classroom/student-level use. This level includes what is usually called classroom assessment, and the primary purpose is to improve instruction in the short run.

In Figure 3, arrows depict the flow of information from the State assessment or the local assessment to one of the three levels of analysis (Notice that the density of the arrows reflects the most likely or heaviest flow of information). The first obvious generalization is that information from either the State assessment or local assessment can be used as raw input for any of the three types of analyses. Our focus here is on the relationships between the sources or origins of the assessment information and its uses. Although any of the three uses can be "fed" by either the State or local source, different patterns carry different obligations for both State and local personnel. Discussion about each of the three levels of analysis follows.

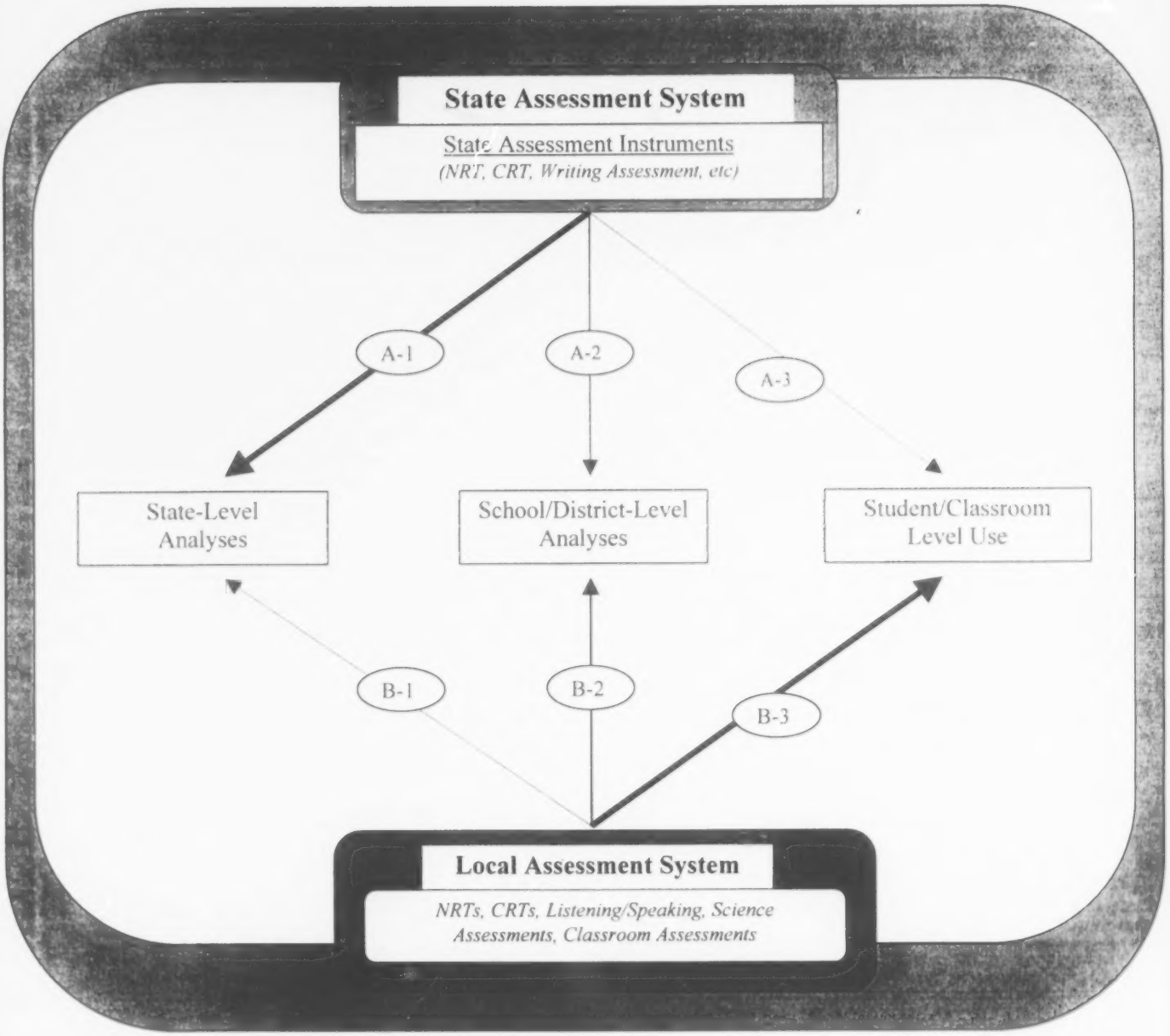


Figure 3

BEST COPY AVAILABLE

State-level analyses

The A-1 arrow, which represents the drawing of statewide inferences about statewide performance from state assessments, is the most routine and obvious. What about B-1? Although far from routine, there are assessment systems in which some standards are assigned to local assessment; the resulting information is reported to the state where it is used to form an overall "score." This is one way that all standards can be assessed as part of the state's assessment system, but not as part of a state assessment, per se. If this approach is taken, it is clear that the state has an obligation both to assist local personnel in administering and scoring student work, and also to monitor the process to ensure objectivity, comparability and accuracy.

School/district-level analyses

The meaning of arrows A-2 and B-2 is quite similar to that of A-1 and B-1 described above. Using the results of a state assessment to judge the progress of learning for a school (A-2) is certainly routine, if not the norm. IASA assumes that the official judgement of school progress would be based on the results of common statewide assessments.

The analysis for B-2 is more complex. It could represent a situation similar to B-1, that is, the official designation of local assessments to assess specific state standards, or aspects of standards. In this case, the obligations for ensuring comparability and accuracy fall heavily upon both state and local agencies.

Other meanings for B-2 exist. For example, the arrow could refer to the use of local assessment information, not necessarily linked directly to State standards, as it is used to confirm or disconfirm the findings from the State assessment, especially as they relate to the identification of schools in need of improvement. In this case, the State assessment is the primary source of data, and the local information is used to document cases where the State results are "in error for statistical or other substantive reasons." In this case, the State's obligation is to set forth the general rules for such disputation, and to monitor their implementation on a statewide basis, perhaps using some kind of sampling or auditing scheme.

Or, as a third meaning, it could simply refer to the LEA exercising its option to assess other aspects of the standards, other standards, or other content areas for the purpose of monitoring individual student progress--none of which would obligate the State to ensure any kind of comparability.

Classroom/student-level use.

Student/classroom level use presents yet another set of options or possibilities. The use of State assessment information at the student level (A-3) can take at least three different forms:

- For high-stakes assessment decisions about individual students, such as promotion or graduation, consequences are usually based on a State assessment that has met rigorous standards of alignment and technical quality.

- In low-stakes situations, many States report results from the statewide assessment for individual students. Because state assessments are administered infrequently and may not reflect local curriculum emphasis, teachers often regard the results as less helpful than classroom assessments to support instructional planning.
- For those State assessment systems that use matrix-sampling to assess the standards at the school level, individual student results may or may not be reported. Some State systems take advantage of the power of matrix-sampling and find a way to report individual results on the whole set of standards. Another alternative is for the State to use a mixture of methods in which the results of a machine-scored assessment are reported at the student level, but performance exercises administered on a matrix-sampling basis are used to judge whether the school as a whole is making progress on the standards that are more difficult to assess.

B-3 is regarded as the most essential type of assessment by many people, especially students, parents and teachers. Classroom assessment is often described as the more student-centered, curriculum-embedded, instructionally-relevant end of the assessment continuum. As such, from the alignment perspective, it focuses on State standards and/or the local "add-ons," and does so in the context of local timing and in ways that are relevant to a given community. Classroom assessments may also include standards that actually could be assessed formally, but are regarded as more appropriately assessed in the classroom context. Unlike B-1 and B-2, however, where scoring, objectivity and comparability are important across classrooms, the emphasis here is on the assessment's usefulness and meaningfulness to the student and the teacher within a particular classroom.

Many States now take responsibility for helping teachers assess and teach the standards by providing prepackaged assessment materials (or combined instructional/assessment materials) that are specifically designed for the classroom setting. The focus of these efforts is usually on the more demanding process of tapping the higher-order thinking skills. From a systems perspective, it is an effective way to remove a troublesome wrinkle in the State-local interface. The irony is that local assessment systems are crucial to real school reform, yet too often the misalignment of local assessments to State standards retards the whole reform process.

Local assessment of State standards

The allocation of some state standards to local assessment results in different types of obligations for a State, depending on the types, purpose, and use of the information. In one way or another, the reasons behind these allocation decisions almost always relate to assessibility. For standards that are difficult to assess in a standardized manner (e.g., ability to use the writing process, capacity to persevere on a multi-stage project, ability to orally defend a piece of work), the local assessment choice is often selected. The State has demonstrated its commitment to those standards but has no practical way to follow through on collection of the data from thousands of students--hence local assessment.

Cost is also an issue in deciding how, and where, a standard should be assessed. For standards that call for broader forms of assessment, States have been known to say it is

too expensive and, therefore, the task of assessing those standards will be assigned to local assessment. This is merely shifting the cost from State to local sources, of course. It is unlikely that the total cost will be lower, and may well be higher. Not only can economies-of-scale apply to State-level assessments, but if the results are used for State level reporting (B-1) or for making school level decisions (B-2), the State has the added cost of the obligation to monitor the accuracy and comparability of the assessments.

Post Script

Most descriptions of assessment systems focus on the grade levels and content areas assessed, and maybe on the source of the items, i.e., the use of ready-made assessments versus customized assessments. But these important elements provide scarcely a clue about the system as a *system* in which the emphasis is on the purposes and the interrelationships among the parts of the system in achieving those purposes. One could even imagine the use of the schematic in Figure 3 as the beginning of a coding system; for example, each of the six data paths/arrows might be coded 0, 1 or 2 denoting the existence of the path in a given program, and the specific type of analysis that it facilitated. Aficionados could have discussions, for example, about the virtues of a 102020 over a 112200 system! More importantly, it could provide a framework for helping public and policy groups (as well as educators at the SEA and LEA level, who might have great expertise in the use of tests, per se, but frequently are as unclear on the different types of assessment systems as the lay public) to have informed discussions about the relative advantages and disadvantages of different configurations.

Appendix C. Summary of Alignment Elements and Illustrative Types and Sources of Evidence

Alignment Element	Illustrative Evidence of Alignment
a. Content Alignment: Comprehensiveness	The State's documentation describes-- • the relationships between the structure of the standards and the structure of the assessments, • the rationale for the overall alignment strategy, including a rationale for any standards either not assessed or reported under the aegis of State assessment, • the manner in which each standard is assessed, in terms of -the three levels of origin and purpose, -the level of assessment formality for each standard, -the type of information the State collects pertaining to each standard, and -how the State monitors the quality of the data collected at the local level, especially if it is part of the statewide Title I system. Considerations include-- *Assessments are described in terms of their purposes, and their roles in the State assessment system.* • The nature of the evidence for alignment is compatible with the assessment's purpose and use, its origin and source--local or State, its level of assessment formality, and the need for technical rigor. *Assessments are designed to match the content standards.* • the assessment blueprint describes how each content standard will be assessed, with appropriate item/task formats for each aspect of the standards, • the blueprint specifies the proportions of the assessment that will cover each content standard, • if domain sampling is used, the blueprint includes methods for ensuring that each domain is adequately covered, and • item/task specifications (for selected-response items and their options and for constructed-response/performance items and their scoring rubrics) specify the ways in which each standard will be assessed.

	All items/tasks are related to the content standards. • each item/task on the assessment measures one or more content standards • for selected-response items, incorrect answers are related to inadequate or incomplete knowledge in the standard(s) assessed • for constructed response and performance items, all criteria in the scoring rubrics are related to the standard(s) assessed • the contexts (e.g., story problems, graphics, texts) for the items/tasks are appropriate to the content standard(s) assessed *The assessment fully covers the content standards.* • all content standards are measured by the assessment, or all important domains of the content area are measured by the assessment • each content standard (or domain) is measured using an appropriate mix of item/task formats
b. Content alignment-Emphasis	Considerations in evaluating "emphasis alignment" include-- • the items/tasks as a whole measure knowledge and skills in a manner that reflects the emphasis of knowledge and skills in the content standards • the formats used to measure different standard(s) reflect the emphasis of types of knowledge and skills in the content standards.
c. Content Alignment-Depth	Assessment blueprint and judgment of experts reflect match between intended depth of assessment (higher order thinking and understanding) and overall depth of assessment. Key considerations include-- • item/task specifications indicate the depth at which knowledge and skills should be measured • each item/task elicits responses reflecting the depth of knowledge and skills in the content standard(s) it measures • each item/task uses an appropriate format for the depth of knowledge and skills in the content standard(s) it measures • as a whole, the assessment reflects the range of depth of knowledge and skills implied by the set of content standards • statistical item/task analyses indicate that items/tasks are at a level of difficulty commensurate with the content standard(s) measured

d. Performance Standards	The assessment tasks match the-- • performance descriptors, and • the exemplars of student work. Considerations in evaluating alignment with the performance standards include-- • the assessment blueprint specifies how the entire range of performance descriptors will be measured by the assessment • item specifications are referenced to the levels of knowledge and skills in the performance descriptors [OR item specifications include guidelines for how items/tasks can measure the levels of knowledge and skills in the performance descriptors] • the assessment as a whole covers knowledge and skills at each defined performance level • each aspect of the performance descriptors is covered by one or more items/tasks • score reports and statistical item/task analyses indicate that students at all performance levels have the opportunity to demonstrate their knowledge and skills • the illustrative student work used to define and communicate the performance levels must match the performance descriptors at each level
e. Clarity and Transparency	Quality and thoroughness of materials used to communicate the alignment between the standards and the assessments, including use of indices, graphics and other devices to communicate the degree of alignment.
Alignment Assurance Process--Who does it and how do they do it?	Criteria relate to the participants, the training, and the materials: Participants. States should consider including panel members with expertise in the following areas: • the content of the standards and assessments • the students to whom the standards and assessments apply • the development and intended use of the content standards, performance standards, and assessment system educational measurement Training. Panelists should be familiar with the following topics (depending upon the composition of the panel, some of these topics may not need to be covered before review): • content standards • performance standards

	• use and purposes of the assessments • student population to which the standards and assessments apply • review process (training should include practice in the process) Materials for Review. Panelists should have access to the following materials during review: • content and performance standards • assessment blueprints • answer keys, scoring rubrics, and scoring guides • assessments • student response information (including sample responses for open-ended items and item/task statistics)

Appendix D.
Summary of Elements and Illustrative Evidence of Technical Quality

Elements of Quality	Illustrative Evidence of Technical Quality
a. Validity **Construct validity** Evidence based on test content (content validity)	Evidence of alignment to content and performance standards (See criteria under "II.A. Alignment," including the six elements and the alignment process.)
Evidence based on student response processes (and evidence based on the test development process)	Assessment development strategy based on • content-process analysis of content standards • theoretical framework linking assessment purposes, parameters of instruments, and reporting strategies Item/task/test development logic and evidence: • characteristics of items, including clarity, readability, etc. • appropriateness of assessment strategy to concept being assessed • evidence of students drawing upon learning from instruction to give high quality responses (test sensitivity to instruction) • evidence based upon use of cognitive lab approach using student think-aloud and other techniques • evidence of level of challenge appropriateness for students at different levels of performance • depth and breadth of item/task review at different stages by appropriate specialists and practitioners • all of above, to ensure appropriateness for ALL students
Evidence based on the assessment's relationship with other variables.	Evidence of confirming relationships: • higher assessment scores related to -teacher judgment of student work meeting standards, including classroom assessment -other school and district administered assessments

	• positive changes in results associated with judged high quality teaching and learning • difference in results between groups exposed to different levels of instruction and practice • greater success in later grades and subjects for students who did well on the assessments Evidence of disconfirming relationships: • lower relationships between assessment results of content standards in question and assessments in other content areas
Evidence based on internal structure (and evidence based on the test development process)	Evidence from • the assessment development process, including results of item analyses, item response scaling, and generalizability studies • special studies using other techniques such as confirmatory factor analysis
Consequential aspects of validity	Evidence of process in place to systematically gather evidence of • greater student learning • better teaching • better teaching and learning for ALL students • more or better staff development opportunities • increase or decrease in breadth of instruction • more or less time devoted to test prep • better use of local and classroom assessment • greater public involvement or support for public schools
b. Reliability	Evidence of identifying and reporting of different sources of variability at • student and school levels, including that associated with test forms, school-by-form interactions, item-person interactions and other relevant types Evidence of • sufficient number of items per student to provide reliable scores at individual level, or for school-level reporting • sufficient opportunity for students to be able to respond at functional level, not guessing or pure knowledge

	• efforts to use this information on sources of variance in the assessment design • sufficient accuracy for reporting of results for disaggregated groups • concordance between level of specificity in reporting results at standard or domain level and likely threats to reliability • sufficient rater agreement for constructed response and extended-response items • sufficient stability to permit study of changes over time—and reporting of same • sufficient homogeneity of items/tasks within a reporting area • school level stability studies conducted to provide evidence of high probability of accurately detecting a school in need of improvement
c. Fairness/Accessibility	State's efforts in attending to four aspects of fairness: 1. The items or tasks provide an equal opportunity for all students to fully demonstrate their knowledge and skills, e.g., they are accessible. Assessments allow for-- • Different ways of expressing competency and responding to tasks, including use of multiple measures to allow all students the opportunity to demonstrate what they know and can do, • The use of accommodations and modifications, • The screening for bias and irrelevant factors, and • The empirical study of items and tasks. Considerations in evaluating accessibility include-- • Groups of selected response items cover a variety of ways of expressing knowledge and skills related to content standard(s) • Constructed response/performance tasks allow for a range of responses to be referenced to each score point in their scoring rubrics • Sample student responses for constructed response/performance tasks contain a full range of response types and levels • Accommodations and modifications are available for students with disabilities, English language learners, and other students who need them in order to demonstrate their level of proficiency in the content area • Items and tasks are appropriate for the age and grade level of the students assessed

- Items/tasks and the assessment as a whole have been reviewed for potential bias against or stereotypical/offensive content about groups of students, including regional populations [doesn't sound right]; students with and without disabilities; racial, ethnic, language, and cultural groups; boys and girls; religious groups; etc.
- The assessment is free of irrelevant factors that are likely to interfere with students' opportunity to demonstrate their knowledge and skills, such as assumptions about background experiences and extraneous prior knowledge
- Statistical item/task analyses (including bias analyses) indicate that all students have the opportunity to demonstrate their knowledge and skills

Some main steps--

- involve item writers and reviewers who are familiar with the different backgrounds and characteristics of these student groups,
- conduct special field testing of all items with all affected student groups, and
- conduct studies, as part of the development process, that identify any items and tasks that might put students with different backgrounds at a disadvantage.

2. The assessments are administered in ways that ensure fairness.

States are responsible for ensuring fairness in the administration of the assessments, including--

- the use of accommodations to allow all students the opportunity to demonstrate their skills,
- the use of assessments in the students' home languages,
- the use of alternate assessments with students whose disabilities prevent them from taking the "regular" assessments, even with accommodations,
- the development and dissemination of clear policies about the use of these variations, and
- the collection and analysis of information not only regarding frequency of use of these variations, but also the effectiveness and the difficulties associated with their use.

3. The results are reported in ways that ensure fairness.

States can help ensure fair reporting of the results by--

- working with LEAs to help them in the proper reporting of results for individual

	students, • designing school-level reports that accurately portray group differences, and • not reporting results for very small groups of students in its disaggregated reporting, in order to ensure confidentiality, and to avoid misinterpretations. 4. The results are interpreted or used in ways that lead to equal treatment. • States can help LEAs and the press to be more careful in interpreting results for all students, and especially for groups of students for whom the assessment may not have yielded accurate results. • States need to develop a long-range strategy that-- • monitors the occurrence and likely evidence of various kinds of improper interpretation and use of the assessment results, and • provides professional development on a statewide basis to help teachers and administrators properly understand the assessment results and how they relate to instruction, especially for students with different backgrounds or with disabilities.
d. Comparability	Evidence that scores have the same meaning from year to year: • administration, scoring and analysis procedures are unchanged or have evidence of changes are of no effect (including context effect due to changes of form configuration) • item content and focus maintained, even if new items are added • level of difficulty maintained or analytic procedures remove likelihood of • misinterpretations of school, LEA and state changes in student progress Evidence that scores have the same meaning from unit to unit: • -schools take same assessment, or • -State provides for fair comparisons by ensuring that different assessments are-- -aligned to standards, and -have essentially identical statistical characteristics, and are -reported on same scale, and have passed other stringent tests of comparability
e. Administration, scoring, analysis, and reporting	

Administration	• Clear and complete directions exist for administering assessments, including procedures for various contingencies • State verifies that assessments are administered according to directions (confirmed via sampling/auditing) • State verifies that security procedures are followed: -materials not accessible except during administration -all materials collected and stored properly • State provides sufficient illustrative and practice materials to give -students adequate practice in marking their answers -teachers illustrative materials samples developing lessons which teach the general skills--but not unduly specific skills • State has clear policies regarding appropriate test preparation practices and verifies that they are followed

Scoring	• State monitors accuracy of machine scoring: scoring keys, scanner accuracy, production of scoring tape/record. • State monitors accuracy and objectivity of human judgment-based scoring: image-based adequacy, clarity of rubrics, training and monitoring of raters.
Analysis	State monitors quality of analysis plans, appropriateness of analyses to purposes, accuracy and fidelity in carrying out the plans, thoroughness of quality control plan
Reporting	State shapes the content, format and timing of the reports to the needs of the users, including: • the timeliness of the reports, • the sufficiency of the descriptive materials, • the level of specificity of the findings, and • the effectiveness of the graphic and other visual devices to communicate either overall findings or relative strengths and weaknesses of certain groups of students or of their performance on different standards. State provides estimates of measurement error that are clearly presented and easily understood to help users make appropriate inferences: • school and individual error are clearly differentiated • includes magnitude of error in terms of probability of misclassification -percentage above cut (PAC) -for each year and for annual changes -for differences among student groups--and for change for various subgroups • all of above for totals and for any content categories reported, whether at standard level or in greater detail Evidence of plans to gather information on effectiveness of different reports with different user groups
f. Interpretation and Use	Evidence that the State has made an effort to assist local and State personnel in proper interpretation and use of results: • staff development focusing on understanding content and purpose of the assessment system • worked with professional associations as allies in staff development

	• efforts to evaluate the effectiveness of State and local efforts • focus of staff development on the role of uncertainty in the interpretation and use of results State takes active role, and has research program in the use of school level results to identify schools for program improvement
Opportunities to gather evidence on technical qualities using three time periods:	
Design/development phase	Emphasis on alignment, and design features especially related to validity, reliability, and fairness.
Initial implementation phase	State confirms the technical characteristics of the assessment tasks, including the fairness of the tasks for different student populations, and confirms the link between the assessments and the performance standards. States should also describe steps taken to confirm proper administration, scoring, analytic procedures and reporting practices.
On-going implementation	This is the time to focus on the proper use and interpretation of the results, and to document consequences of assessment.